Tractable Learning of Graphical Model Structures from Data

A Dissertation Presented

by

Jean Fausto Honorio Carrillo

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Computer Science

Stony Brook University

August 2012

Stony Brook University

The Graduate School

Jean Fausto Honorio Carrillo

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Dimitris Samaras - Dissertation Advisor
Associate Professor, Computer Science

Tamara Berg - Chairperson of Defense
Assistant Professor, Computer Science

Luis Ortiz - Dissertation Co-advisor
Assistant Professor, Computer Science

Tommi Jaakkola - Outside Member
Professor, Electrical Engineering and Computer Science, Massachusetts Institute of Technology

This dissertation is accepted by the Graduate School

Charles Taber
Interim Dean of the Graduate School

Abstract of the Dissertation

Tractable Learning of Graphical Model Structures from Data

by

Jean Fausto Honorio Carrillo

Doctor of Philosophy

in

Computer Science

Stony Brook University

2012

Probabilistic graphical models (PGMs) provide a way to represent variables (nodes) along with their conditional dependencies (edges) and therefore allow formalizing our knowledge of the interacting entities in the real world. Structure learning aims to discover the topology (and parameters) of a PGM that represents accurately a given dataset. Accuracy of representation can be measured by the likelihood that the PGM explains the observed data, which leads to maximum likelihood estimation (MLE).

From an algorithmic point of view, one challenge faced by structure learning is that the number of possible structures is super-exponential in the number of variables. From a statistical perspective, it is important to find good regularizers in order to avoid over-fitting and to achieve better generalization performance. Regularizers aim to reduce the complexity of the PGM, which can be measured by its number of parameters.

First, we present three regularizers for MLE of Gaussian Markov random fields (MRFs): local constancy for datasets where variables correspond to a measurement in a manifold (silhouettes, motion trajectories, 2D and 3D images); variable selection for finding few interacting nodes from datasets with thousands of variables; and multi-task learning for a more efficient use of data which is available for multiple related tasks. For these regularizers, we show bounds of the eigenvalues of the optimal solution, convergence of block coordinate descent optimization, and connections to the continuous quadratic knapsack problem and the quadratic trust-region problem.

Second, we focus on learning sparse discrete MRFs through MLE. In this case, computing the objective function as well as its gradient is NP-hard. We study the convergence rate of stochastic optimization of exact NP-hard objectives, for which only biased estimates of the gradient are available. We provide a convergence-rate analysis of deterministic errors and extend our analysis to biased stochastic errors.

Third, we show general results for PGMs that allow understanding MLE with regularizers on the differences of parameters (e.g. sparse structural changes, time-varying models), the generalization ability of PGMs, and the use of PGM parameters as features in classification, dimensionality reduction and clustering. To this end, we show that the log-likelihood of several PGMs is Lipschitz continuous with respect to the parameters, and derive bounds on the Kullback-Leibler divergence, expected log-likelihood and Bayes error rate.

Finally, we formalize and study the problem of learning the structure of graphical games from strictly behavioral data. We propose MLE of a generative model defined by the Nash equilibria of the game. The formulation brings out the interplay between goodness-of-fit and model complexity: good models capture the equilibrium behavior represented in the data while controlling the true number of equilibria, including those potentially unobserved. We provide a generalization bound for MLE. We discuss several optimization algorithms including convex loss minimization, sigmoidal approximations and exhaustive search. We formally prove that games in our hypothesis space have a small true number of equilibria, with high probability; thus, convex loss minimization is sound.

We present experimental results on a wide range of real-world datasets: walking video sequences, motion capture, cardiac MRI, brain fMRI, gene expression, stock prices, world weather and congressional voting.

Regarding published work, under submission and soon to be submitted. Chapter 2 is based on published work [Honorio et al., 2009], additionally it includes bounds on the eigenvalues of the optimal solution. Chapter 3 is based on published work [Honorio et al., 2012]. Chapter 4 is based on published work [Honorio and Samaras, 2010] for the $\ell_{1,\infty}$ penalty, and work under submission [Honorio and Samaras, 2012] for the $\ell_{1,2}$ penalty and diagonal penalization. Chapter 6 is based on published work [Honorio, 2011]. Chapter 5 is based on published work [Honorio, 2012]. Chapter 7 is based on work soon to be submitted [Honorio and Ortiz, 2012].

Dedication Page

I dedicate this thesis to my fiancée Verónica Peña for bringing happiness into my life, for her unconditional support, for her patience, for her understanding, for all the minutes that I stole from us, for all the hours that I will never be able to return, for making the biggest sacrifice, for always remembering what is important in our lives, for reminding it to me with her smile.

I dedicate this work to my friends Chihiro Suzuki, Siti Rokhmah, Fumito Hiraoka, Tejdipto Bose and Waqar Ahmad, for keeping in touch besides me living in seclusion from society.

Last but not least, I dedicate this thesis to my parents Jaime and Mercedes, and to my sisters Karen and Josseline, for providing me with an infinite source of encouragement.

Contents

List of Figures

List of Tables

List of Algorithms

Acknowledgements

I would like to express my sincere gratitude to my thesis advisor Prof. Dimitris Samaras and my co-advisor Prof. Luis Ortiz, for their trust and support, for the privilege of being able to work on topics of mutual interest, and for giving me freedom for pursuing topics of my interest.

I am greatly indebted to my thesis committee members: Prof. Tamara Berg and Prof. Tommi Jaakkola. Their expert, constructive and fruitful comments have been very valuable for this thesis, and provide me inspiration for my future research.

I would like to thank my collaborators at the Brookhaven National Laboratory: Rita Goldstein, Dardo Tomasi, Nelly Klein, Thomas Maloney, Scott Moeller and Anna Konova, as well as Prof. Hoi-Chung Leung from the Psychology Department. Our discussions on neuro-scientific problems have been a source of inspiration for part of this work. Additionally, I would like to thank Rita Goldstein for providing us the cocaine addiction fMRI dataset, and Dardo Tomasi for providing us the *1000 functional connectomes* dataset.

I would also like to thank my colleagues in École Centrale Paris for having me invited to visit their group: Prof. Nikos Paragios, Prof. Iasonas Kokkinos, Aristeidis Sotiras, Ahmed Besbes, Loïc Simon, Olivier Teboul, Salma Essafi and Katerina Gkirtzou. Many thanks go to Chaohui Wang for being such a good host, to Martin de la Gorce for our few but very interesting technical discussions in the office and in the train, and to Georg Langs from the University of Vienna, who I worked with during my visit to École Centrale Paris.

I am very grateful to my collaborators at IBM Thomas J. Watson Research Center: Irina Rish and Guillermo Cecchi. We share several common interests and our short interactions have proven to be very fruitful.

I would like to express my sincere gratitude to all my colleagues and friends from the Image Analysis lab: Prof. Alexander Berg, Alexandros Panagopoulos, Yun Zeng (a.k.a. Xiang Zeng), Wei Hu, Tejo Chalasani, Hojin Choi, Xufeng Han, Kiwon Yun, Yifan Peng, Tomás Yago, Chen-Ping Yu, Lei Fan, Vicente Ordoñez and Kota Yamaguchi. I would also like to thank Mohammad Irfan and Hau Chan from the Machine Learning lab.

I am very grateful to Nicolle Gruzling for sharing her Master's thesis.

This work was done while the author was supported in part by NIH Grants 1 R01 DA020949, 1 R01 DA023579, 1 R01 EB007530 and NSF Grant CNS-0721701.

Chapter 1

Introduction

In this chapter, we provide a brief introduction to probabilistic graphical models and the problems of inference and learning. Since we propose different priors for learning Gaussian graphical models in Chapters 2, 3 and 4, here we provide the material that is common to these chapters. We provide the material for discrete Markov random fields in Chapter 5.

1.1 Probabilistic Graphical Models

The availability of large amounts of real world data and the need of not only being able to detect patterns but also to understand its underlying structure has pushed the development of graphical models in the last years. When modeling real world data, a probabilistic model must deal with uncertainty and complexity. Graphical models provide a way to represent variables along with their conditional dependencies and therefore allow formalizing our knowledge of the interacting entities in the real world related to the problem at hand. In computer vision, for instance, an image can be decomposed into its constituent objects, lighting conditions, motion patterns, etc. Note however that even though knowledge about the lighting conditions and their effect to the image is known, it is unobserved since it is not usually recorded along with the data.

Inference algorithms must deal with uncertainty coming from different sources: the data, the features which are most useful for processing the data, the relationship among variables, and the action which has to be taken as a result of inference [Frey and Jojic, 2005]. Probability theory offers two types of models when reasoning under uncertainty.

A *discriminative model* predicts the distribution of the desired output given the input data $P(output|input)$. In tasks such as supervised learning the output is the class label and the likelihood $P(input|output)$ is usually learnt. The posterior probability is computed by the Bayes rule. On the other hand, a *generative model* allows modeling the data as a joint probability distribution $P(output, input)$. The posterior probability $P(output|input)$ is computed by marginalization and the Bayes rule.

However, sometimes our goal is not only to find a model which fits the training data, but also that is consistent with prior knowledge. *Graphical models* provide a way to specify prior knowledge about the variables and their topology (conditional dependencies). In the presence of knowledge about the process which generated the data, they also allow defining

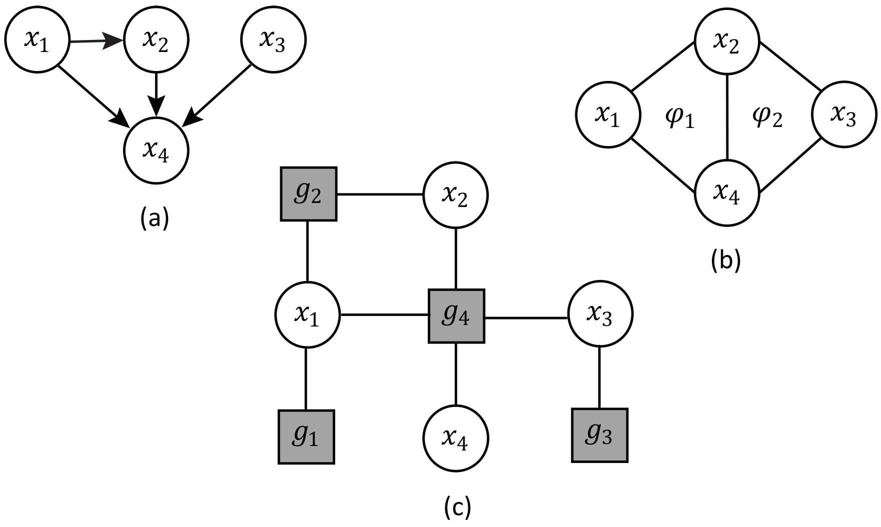


Figure 1.1: Three types of graphical models representing the same conditional dependencies, (a) a Bayesian network, (b) a Markov random field and (c) a factor graph.

hidden or unobserved variables which explain the observed data. This results in an extended joint distribution of the form $P(class, input, hidden)$.

1.2 Types of Graphical Models

Bayesian or belief networks [Bishop, 2006, Frey and Jojic, 2005, Lauritzen, 1996] are directed acyclic graphs where each node is a random variable and each edge represents conditional dependence of a variable with respect to its parent. More formally, a Bayesian network for random variables $x_1, x_2 \ldots x_N$ has one conditional probability function $P(x_n|\phi_n)$ for each variable x_n given its set of parents ϕ_n. The joint probability distribution is given by the product of all conditional probabilities $P(x) = \prod_n P(x_n|\phi_n)$. For instance, in Figure 1.1(a) we show a Bayesian network in which x_4 depends on x_1, x_2 and x_3, while x_2 and x_3 are conditionally independent.

A *Markov random field* [Bishop, 2006, Frey and Jojic, 2005, Lauritzen, 1996] for random variables $x_1, x_2 \ldots x_N$ is an undirected graph with one potential function g_m for each of the M maximal cliques φ_m. The joint distribution is given by the product of all potential functions $P(x) = \frac{1}{Z} \prod_m g_m(\varphi_m)$ where Z is a normalization constant such that $\int P(x)dx = 1$. Figure 1.1(b) shows an example in which two cliques are defined.

Factor graphs [Bishop, 2006, Frey and Jojic, 2005] allow splitting a joint probability distribution of several variables into a product of local functions of smaller sets of variables. More formally, a factor graph is a bipartite graph where one set of nodes are the random variables $x_1, x_2 \ldots x_N$ while the other set are the local functions $g_1, g_2 \ldots g_M$. A function g_m is connected to the set variables of variables φ_m on which it depends on. The joint probability distribution is given by the product of all local functions $P(x) = \frac{1}{Z} \prod_m g_m(\varphi_m)$ where Z is a normalization constant such that $\int P(x)dx = 1$. See Figure 1.1(c) for a graphical example.

Factor graphs subsume Bayesian networks and Markov random fields. Any Bayesian network or Markov random field can be converted to a factor graph. Furthermore, there exist models whose independence relationships can only be expressed in a factor graph.

Markov blankets allow understanding the concept of independence in graphical models. The Markov blanket of a node contains all the variables that shield the node from the rest of the model. This means that the Markov blanket of a node is the only knowledge needed in order to predict the behavior of that node. For Bayesian networks, the Markov blanket of a node includes its parents, children and children's parents. For Markov random fields, the Markov blanket of a node includes all its neighbors, which is the set of maximal cliques the node belongs to. For factor graphs, the Markov blanket of a node includes all its neighbors, which are all local functions connected to it.

1.3 Inference and Learning

Two important tasks related to graphical models are inference and learning. *Inference* refers to the problem of computing the value of a subset of nodes (hidden or unobserved variables) given observed values in another subset (observed variables) [Bishop, 2006]. The hidden values are computed by maximization of the posterior $P(hidden|input)$ which can be computed by the Bayes rule. Exact inference by posterior maximization is computationally intractable in most cases, for instance when some variables are discrete, all possible combinations have to be tried in the search. Due to that, inference is performed by using an approximate and simpler probability density function or by some heuristics. Several approximate inference algorithms have been proposed [Frey and Jojic, 2005] such as iterated conditional modes, expectation maximization, Monte Carlo methods, variational techniques as well as belief propagation.

Parameter learning refers to the problem of computing the value of the parameters for a graphical model with fixed topology. *Structure learning*, the topic of this thesis, refers to the problem of learning the topology (and parameters) of the graphical model. In both cases, the learning process uses a given dataset in order to estimate the parameters and/or structure. Techniques on this area have a big impact in contexts with limited prior knowledge where the existence of variables is known but not the conditional dependencies among them.

Even though some research in structure learning has been done for datasets with missing values on some samples [Myers et al., 1999, Ramoni and Sebastiani, 1997] as well as for hidden or unobserved variables [Elidan and Friedman, 2005, Friedman, 1997, 1998], we focus on datasets without missing values and with observed variables only.

1.4 Learning Sparse Gaussian MRFs

We propose different priors for learning Gaussian graphical models in Chapters 2, 3 and 4. In this section, we introduce Gaussian graphical models and discuss methods for learning sparse models from data.

A *Gaussian graphical model* is a graph in which all random variables are continuous and jointly Gaussian. This model corresponds to the multivariate normal distribution for N variables with covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{N \times N}$. Conditional independence in a Gaussian graphical model is simply reflected in the zero entries of the precision matrix $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ [Lauritzen, 1996]. Let $\mathbf{\Omega} = \{\omega_{n_1 n_2}\}$, two variables n_1 and n_2 are conditionally independent if and only if $\omega_{n_1 n_2} = 0$. The precision matrix representation is preferred because it allows detecting cases in which two seemingly correlated variables, actually depend on a third confounding variable.

For Gaussian graphical models, the number of parameters, the number of edges in the structure and the number of non-zero elements in the inverse covariance or precision matrix are equivalent measures of complexity. Therefore, several techniques focus on enforcing sparseness of the precision matrix.

The concept of robust estimation by performing covariance selection was first introduced in [Dempster, 1972] where the number of parameters to be estimated is reduced by setting some elements of the precision matrix $\mathbf{\Omega}$ to zero. Since finding the most sparse precision matrix which fits a dataset is a NP-hard problem [Banerjee et al., 2006], in order to overcome it, several ℓ_1-regularization methods have been proposed for learning Gaussian graphical models from data.

Given a dense sample covariance matrix $\widehat{\mathbf{\Sigma}} \succeq \mathbf{0}$, the problem of finding a sparse precision matrix $\mathbf{\Omega}$ by regularized maximum likelihood estimation is given by:

$$\max_{\mathbf{\Omega} \succ \mathbf{0}} \Big(\log\det \mathbf{\Omega} - \langle \widehat{\mathbf{\Sigma}}, \mathbf{\Omega} \rangle - \rho \|\mathbf{\Omega}\|_1 \Big) \tag{1.1}$$

for $\rho > 0$. The term $\log\det \mathbf{\Omega} - \langle \widehat{\mathbf{\Sigma}}, \mathbf{\Omega} \rangle$ is the Gaussian log-likelihood. The term $\|\mathbf{\Omega}\|_1$ encourages sparseness of the precision matrix or conditional independence among variables.

Covariance selection [Banerjee et al., 2006] computes small perturbations on the sample covariance matrix such that it generates a sparse precision matrix, which results in a box-constrained quadratic programming. This method has moderate run time. The *Meinshausen-Bühlmann approximation* [Meinshausen and Bühlmann, 2006] obtains the conditional dependencies by performing a sparse linear regression for each variable, by using *lasso* regression [Tibshirani, 1996]. This method is very fast but does not yield good estimates for lightly regularized models, as noted in [Friedman et al., 2007b]. The constrained optimization version of eq.(1.1) is solved in [Yuan and Lin, 2007] by applying a standard determinant maximization with linear inequality constraints, which requires iterative linearization of $\|\mathbf{\Omega}\|_1$. This technique in general does not yield the maximum likelihood estimator, as noted in [Banerjee et al., 2008]. The *graphical lasso* technique [Friedman et al., 2007b] solves the dual form of eq.(1.1), which results in a lasso regression problem. This method has run times comparable to [Meinshausen and Bühlmann, 2006] without sacrificing accuracy in the maximum likelihood estimator.

Structure learning through ℓ_1-regularization has been also proposed for different types of graphical models: Markov random fields (MRFs) by a clique selection heuristic and approximate inference [Lee et al., 2006a]; Bayesian networks on binary variables by logistic regression [Schmidt et al., 2007b]; Conditional random fields by pseudo-likelihood and block regularization in order to penalize all parameters of an edge simultaneously [Schmidt et al., 2008]; and Ising models, i.e. MRFs on binary variables with pairwise interactions, by logistic regression [Wainwright et al., 2006] which is similar in spirit to [Meinshausen and Bühlmann, 2006].

Table 1.1: Notation used in this thesis.

Notation	**Description**
$\lVert\mathbf{c}\rVert_1$	ℓ_1-norm of $\mathbf{c} \in \mathbb{R}^N$, i.e. $\sum_n \lvert c_n\rvert$
$\lVert\mathbf{c}\rVert_\infty$	ℓ_∞-norm of $\mathbf{c} \in \mathbb{R}^N$, i.e. $\max_n \lvert c_n\rvert$
$\lVert\mathbf{c}\rVert_2$	Euclidean norm of $\mathbf{c} \in \mathbb{R}^N$, i.e. $\sqrt{\sum_n c_n^2}$
$\mathbf{Diag}(\mathbf{c}) \in \mathbb{R}^{N\times N}$	matrix with elements of $\mathbf{c} \in \mathbb{R}^N$ on its diagonal
$\mathbf{A} \succeq \mathbf{0}$	$\mathbf{A} \in \mathbb{R}^{N\times N}$ is symmetric and positive semidefinite
$\mathbf{A} \succ \mathbf{0}$	$\mathbf{A} \in \mathbb{R}^{N\times N}$ is symmetric and positive definite
$\lVert\mathbf{A}\rVert_1$	ℓ_1-norm of $\mathbf{A} \in \mathbb{R}^{M\times N}$, i.e. $\sum_{mn} \lvert a_{mn}\rvert$
$\lVert\mathbf{A}\rVert_\infty$	ℓ_∞-norm of $\mathbf{A} \in \mathbb{R}^{M\times N}$, i.e. $\max_{mn} \lvert a_{mn}\rvert$
$\lVert\mathbf{A}\rVert_2$	spectral norm of $\mathbf{A} \in \mathbb{R}^{N\times N}$, i.e. the maximum eigenvalue of $\mathbf{A} \succ \mathbf{0}$
$\lVert\mathbf{A}\rVert_{\mathfrak{F}}$	Frobenius norm of $\mathbf{A} \in \mathbb{R}^{M\times N}$, i.e. $\sqrt{\sum_{mn} a_{mn}^2}$
$\langle\mathbf{A}, \mathbf{B}\rangle$	scalar product of $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{M\times N}$, i.e. $\sum_{mn} a_{mn} b_{mn}$
$\mathbf{A} \circ \mathbf{B} \in \mathbb{R}^{M\times N}$	Hadamard or entrywise product of $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{M\times N}$, i.e. $(\mathbf{A} \circ \mathbf{B})_{mn} = a_{mn} b_{mn}$
$\mathbf{Diag}(\mathbf{A}) \in \mathbb{R}^{N\times N}$	matrix with diagonal elements of $\mathbf{A} \in \mathbb{R}^{N\times N}$ only
$\mathbf{diag}(\mathbf{A}) \in \mathbb{R}^N$	vector with diagonal elements of $\mathbf{A} \in \mathbb{R}^{N\times N}$
$\mathbf{vec}(\mathbf{A}) \in \mathbb{R}^{MN}$	vector containing all elements of $\mathbf{A} \in \mathbb{R}^{M\times N}$
$\partial f/\partial\mathbf{c}$	gradient of f with respect to $\mathbf{c} \in \mathbb{R}^N$, i.e. $\partial f/\partial\mathbf{c} \in \mathbb{R}^N$
$\partial f/\partial\mathbf{A}$	gradient of f with respect to $\mathbf{A} \in \mathbb{R}^{M\times N}$, i.e. $\partial f/\partial\mathbf{A} \in \mathbb{R}^{M\times N}$
$\lvert\mathbf{c}\rvert$	entrywise absolute value of $\mathbf{c} \in \mathbb{R}^N$, i.e. $(\lvert c_1\rvert, \lvert c_2\rvert, \ldots, \lvert c_N\rvert)^{\mathrm{T}}$
$\mathbf{J}(\mathbf{A}) \in \mathbb{R}^{M\times N}$	*zero structure operator* of $\mathbf{A} \in \mathbb{R}^{M\times N}$, by using the Iverson bracket $j_{mn}(\mathbf{A}) = 1[a_{mn} = 0]$
$\mathbf{A} \oslash \mathbf{B} \in \mathbb{R}^{M\times N}$	*diagonal excluded product* of $\mathbf{A} \in \mathbb{R}^{M\times N}$ and $\mathbf{B} \in \mathbb{R}^{N\times N}$, i.e. $\mathbf{A} \oslash \mathbf{B} = \mathbf{J}(\mathbf{A}) \circ (\mathbf{A}\mathbf{B})$. It has the property that no diagonal entry of $\mathbf{B}$ is used in $\mathbf{A} \oslash \mathbf{B}$

1.5 Notation

In this thesis, we use the notation in Table 1.1. For convenience, we define two new operators that are used on Chapter 2: the zero structure operator and the diagonal excluded product.

Chapter 2

Learning Independent Gaussian MRFs: Local Constancy

Locality information is crucial in datasets where each variable corresponds to a measurement in a manifold (silhouettes, motion trajectories, 2D and 3D images). Although these datasets are typically under-sampled and high-dimensional, they often need to be represented with low-complexity statistical models, which are comprised of only the important probabilistic dependencies in the datasets. Most methods attempt to reduce model complexity by enforcing structure sparseness. However, sparseness cannot describe inherent regularities in the structure. Hence, in this chapter we first propose a new prior for Gaussian graphical models which, together with sparseness, imposes local constancy. Second, we propose an efficient algorithm which decomposes the strictly convex maximum likelihood estimation into a sequence of problems with closed form solutions. We test our method in a wide range of complex real-world datasets and demonstrate that it captures useful structures such as the rotation and shrinking of a beating heart, motion correlations between body parts during walking and functional interactions of brain regions.

2.1 Introduction

In this chapter, we propose *local constancy* as a prior for learning Gaussian graphical models, which is natural for spatial datasets such as those encountered in computer vision [Crandall et al., 2005, Felzenszwalb and Huttenlocher, 2005, Gu et al., 2007].

In datasets which are a collection of measurements for variables with some spatial arrangement, one can define a local neighborhood for each variable or manifold. Such variables correspond to points in silhouettes, pixels in 2D images or voxels in 3D images. Silhouettes define a natural one-dimensional neighborhood in which each point has two neighbors on each side of the closed contour. Similarly, one can define a four-pixel neighborhood for 2D images as well as six-pixel neighborhood for 3D images. However, there is little research on spatial regularization for structure learning. Some methods assume a one-dimensional spatial neighborhood (e.g. silhouettes) and that variables far apart are only weakly correlated [Levina et al., 2008], interaction between a priori known groups of variables as in [Duchi et al., 2008a], or block structures as in [Mansinghka et al., 2006] in the context of Bayesian networks.

Our contribution in this chapter is two-fold. First, we propose *local constancy*, which encourages finding connectivities between two close or distant clusters of variables, instead of between isolated variables. It does not heavily constrain the set of possible structures, since it only imposes restrictions of spatial closeness for each cluster independently, but not between clusters. We impose an ℓ_1-norm penalty for differences of spatially neighboring variables, which allows obtaining locally constant models that preserve sparseness, unlike ℓ_2-norm penalties. Our model is strictly convex and therefore has a global minimum. Positive definiteness of the estimated precision matrix is also guaranteed, since this is a necessary condition for the definition of a multivariate normal distribution.

Second, since optimization methods for structure learning on Gaussian graphical models [Banerjee et al., 2006, Friedman et al., 2007b, Meinshausen and Bühlmann, 2006, Yuan and Lin, 2007] are unable to handle local constancy constraints, we propose an efficient algorithm by maximizing with respect to one row and column of the precision matrix at a time. By taking directions involving either one variable or two spatially neighboring variables, the problem reduces to minimization of a piecewise quadratic function, which can be performed in closed form.

We initially test the ability of our method to recover the ground truth structure from data, of a complex synthetic model which includes locally and not locally constant interactions as well as independent variables. Our method outperforms the state-of-the-art structure learning techniques [Banerjee et al., 2006, Friedman et al., 2007b, Meinshausen and Bühlmann, 2006] for datasets with both small and large number of samples. We further show that our method has better generalization performance on real-world datasets. We demonstrate the ability of our method to discover useful structures from datasets with a diverse nature of probabilistic relationships and spatial neighborhoods: manually labeled silhouettes in a walking sequence, cardiac magnetic resonance images (MRI) and functional brain MRI.

Section 2.2 introduces techniques for learning Gaussian graphical models from data. Section 2.3 presents our sparse and locally constant Gaussian graphical models. Section 2.4 describes our structure learning algorithm. Experimental results on synthetic and real-world datasets are shown and explained in Section 2.5. Main contributions and results are summarized in Section 2.6.

2.2 Background

In Section 1.4, we introduced Gaussian graphical models as well as techniques for learning sparse Gaussian graphical models through ℓ_1 regularization, such as: *covariance selection* [Banerjee et al., 2006], *graphical lasso* [Friedman et al., 2007b] and the *Meinshausen-Bühlmann approximation* [Meinshausen and Bühlmann, 2006].

There is little work on spatial regularization for structure learning. Adaptive banding on the Cholesky factors of the precision matrix has been proposed in [Levina et al., 2008]. Instead of using the traditional lasso penalty, a nested lasso penalty is enforced. Entries at the right end of each row are promoted to zero faster than entries close to the diagonal. The main drawback of this technique is the assumption that the more far apart two variables are the more likely they are to be independent. Grouping of entries in the precision matrix into disjoint subsets has been proposed in [Duchi et al., 2008a]. Such subsets can model for instance dependencies between different groups of variables in the case of block structures. Although such a formulation allows for more general settings, its main disadvantage is the need for an a priori segmentation of the entries in the precision matrix.

Related approaches have been proposed for Bayesian networks. In [Mansinghka et al., 2006] it is assumed that variables belong to unknown classes and probabilities of having edges among different classes were enforced to account for structure regularity, thus producing block structures only.

2.3 Sparse and Locally Constant Gaussian Graphical Models

First, we describe our local constancy assumption and its use to model the spatial coherence of dependence/independence relationships. *Local constancy* is defined as follows: if variable x_{n_1} is dependent (or independent) of variable x_{n_2}, then a spatial neighbor $x_{n'_1}$ of x_{n_1} is more likely to be dependent (or independent) of x_{n_2}. This encourages finding connectivities between two close or distant clusters of variables, instead of between isolated variables. Note that local constancy imposes restrictions of spatial closeness for each cluster independently, but not between clusters.

In this chapter, we impose constraints on the difference of entries in the precision matrix $\mathbf{\Omega} \in \mathbb{R}^{N \times N}$ for N variables, which correspond to spatially neighboring variables. Let $\widehat{\mathbf{\Sigma}} \in \mathbb{R}^{N \times N}$ be the dense sample covariance matrix and $\mathbf{D} \in \mathbb{R}^{M \times N}$ be the discrete derivative operator on the manifold, where $M \in O(N)$ is the number of spatial neighborhood relationships. For instance, in a 2D image, M is the number of pixel pairs that are spatial neighbors on the manifold. More specifically, if pixel n_1 and pixel n_2 are spatial neighbors, we include a row m in $\mathbf{D}$ such that $d_{mn_1} = 1$, $d_{mn_2} = -1$ and $d_{mn_3} = 0$ for $n_3 \notin \{n_1, n_2\}$. The following penalized maximum likelihood estimation is proposed:

$$\max_{\mathbf{\Omega} \succ \mathbf{0}} \Big(\log\det \mathbf{\Omega} - \langle \widehat{\mathbf{\Sigma}}, \mathbf{\Omega} \rangle - \rho \|\mathbf{\Omega}\|_1 - \tau \|\mathbf{D} \oslash \mathbf{\Omega}\|_1 \Big) \tag{2.1}$$

for some $\rho, \tau > 0$. The first two terms model the quality of the fit of the estimated multivariate normal distribution to the dataset. The third term $\rho\|\mathbf{\Omega}\|_1$ encourages sparseness while the fourth term $\tau\|\mathbf{D} \oslash \mathbf{\Omega}\|_1$ encourages local constancy in the precision matrix by penalizing the differences of spatially neighboring variables.

In conjunction with the ℓ_1-norm penalty for sparseness, we introduce an ℓ_1-norm penalty for local constancy. As discussed further in [Tibshirani et al., 2005], ℓ_1-norm penalties lead to locally constant models which preserve sparseness, where as ℓ_2-norm penalties of differences fail to do so.

The use of the diagonal excluded product for penalizing differences instead of the regular product of matrices, is crucial. The regular product of matrices would penalize the difference between the diagonal and off-diagonal entries of the precision matrix, and potentially destroy positive definiteness of the solution for strongly regularized models.

In the following theorem, we show that the eigenvalues of the optimal solution of our problem is bounded.

Theorem 2.1. *For $\rho > 0$, $\tau > 0$, the optimal solution to the variable-selection structure learning problem in eq.*(2.1) *is unique and bounded as follows:*

$$\left(\frac{1}{\|\widehat{\boldsymbol{\Sigma}}\|_2 + N\rho + \sqrt{MN\min(M,N)}\tau\|\mathbf{D}\|_{\mathfrak{F}}}\right)\mathbf{I} \preceq \boldsymbol{\Omega}^* \preceq \left(\frac{N}{\rho}\right)\mathbf{I} \tag{2.2}$$

Proof. By using the following identities for dual norms $\rho\|\boldsymbol{\Omega}\|_1 = \max_{\|\mathbf{A}\|_\infty \leq \rho} \langle \mathbf{A}, \boldsymbol{\Omega}\rangle$ and $\tau\|\mathbf{D} \oslash \boldsymbol{\Omega}\|_1 = \max_{\|\mathbf{B}\|_\infty \leq \tau} \langle \mathbf{B}, \mathbf{D} \oslash \boldsymbol{\Omega}\rangle$, and the fact that $\langle \mathbf{B}, \mathbf{D} \oslash \boldsymbol{\Omega}\rangle = \langle \mathbf{B}, \mathbf{J}(\mathbf{D}) \circ (\mathbf{D}\boldsymbol{\Omega})\rangle = \langle \mathbf{B} \circ \mathbf{J}(\mathbf{D}), \mathbf{D}\boldsymbol{\Omega}\rangle = \langle \mathbf{D}^{\mathrm{T}}(\mathbf{B} \circ \mathbf{J}(\mathbf{D})), \boldsymbol{\Omega}\rangle$ in eq.(2.1), we get:

$$\max_{\boldsymbol{\Omega} \succ \mathbf{0}} \min_{\substack{\|\mathbf{A}\|_\infty \leq \rho \\ \|\mathbf{B}\|_\infty \leq \tau}} \Big(\log\det\boldsymbol{\Omega} - \langle \widehat{\boldsymbol{\Sigma}} + \mathbf{A} + \mathbf{D}^{\mathrm{T}}(\mathbf{B} \circ \mathbf{J}(\mathbf{D})), \boldsymbol{\Omega}\rangle\Big) \tag{2.3}$$

By virtue of Sion's minimax theorem, we can swap the order of max and min. Furthermore, note that the optimal solution of the inner equation is given by $\boldsymbol{\Omega} = (\widehat{\boldsymbol{\Sigma}} + \mathbf{A} + \mathbf{D}^{\mathrm{T}}(\mathbf{B} \circ \mathbf{J}(\mathbf{D})))^{-1}$. By replacing this solution in eq.(2.3), we get the dual problem of eq.(2.1):

$$\min_{\substack{\|\mathbf{A}\|_\infty \leq \rho \\ \|\mathbf{B}\|_\infty \leq \tau}} \Big(-\log\det(\widehat{\boldsymbol{\Sigma}} + \mathbf{A} + \mathbf{D}^{\mathrm{T}}(\mathbf{B} \circ \mathbf{J}(\mathbf{D}))) - N\Big) \tag{2.4}$$

In order to find a lower bound for the minimum eigenvalue of $\boldsymbol{\Omega}^*$, note that $\|\boldsymbol{\Omega}^{*-1}\|_2 = \|\widehat{\boldsymbol{\Sigma}} + \mathbf{A} + \mathbf{D}^{\mathrm{T}}(\mathbf{B} \circ \mathbf{J}(\mathbf{D}))\|_2 \leq \|\widehat{\boldsymbol{\Sigma}}\|_2 + \|\mathbf{A}\|_2 + \|\mathbf{D}^{\mathrm{T}}(\mathbf{B} \circ \mathbf{J}(\mathbf{D}))\|_2 \leq \|\widehat{\boldsymbol{\Sigma}}\|_2 + N\|\mathbf{A}\|_\infty + \|\mathbf{D}^{\mathrm{T}}(\mathbf{B} \circ \mathbf{J}(\mathbf{D}))\|_{\mathfrak{F}} \leq \|\widehat{\boldsymbol{\Sigma}}\|_2 + N\|\mathbf{A}\|_\infty + \|\mathbf{D}\|_{\mathfrak{F}}\|\mathbf{B} \circ \mathbf{J}(\mathbf{D})\|_{\mathfrak{F}} \leq \|\widehat{\boldsymbol{\Sigma}}\|_2 + N\|\mathbf{A}\|_\infty + \|\mathbf{D}\|_{\mathfrak{F}}\|\mathbf{B}\|_{\mathfrak{F}} \leq \|\widehat{\boldsymbol{\Sigma}}\|_2 + N\|\mathbf{A}\|_\infty + \sqrt{\min(M,N)}\|\mathbf{D}\|_{\mathfrak{F}}\|\mathbf{B}\|_2 \leq \|\widehat{\boldsymbol{\Sigma}}\|_2 + N\|\mathbf{A}\|_\infty + \sqrt{MN\min(M,N)}\|\mathbf{D}\|_{\mathfrak{F}}\|\mathbf{B}\|_\infty \leq \|\widehat{\boldsymbol{\Sigma}}\|_2 + N\rho + \sqrt{MN\min(M,N)}\tau\|\mathbf{D}\|_{\mathfrak{F}}$.

In order to find an upper bound for the maximum eigenvalue of $\boldsymbol{\Omega}^*$, note that, at optimum, the primal-dual gap is zero:

$$-N + \langle \widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Omega}^*\rangle + \rho\|\boldsymbol{\Omega}^*\|_1 + \tau\|\mathbf{D} \oslash \boldsymbol{\Omega}^*\|_1 = 0 \tag{2.5}$$

The upper bound is found as follows: $\|\boldsymbol{\Omega}^*\|_2 \leq \|\boldsymbol{\Omega}^*\|_{\mathfrak{F}} \leq \|\boldsymbol{\Omega}^*\|_1 = (N - \langle \widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Omega}^*\rangle - \tau\|\mathbf{D} \oslash \boldsymbol{\Omega}^*\|_1)/\rho$. Note that $\tau\|\mathbf{D} \oslash \boldsymbol{\Omega}^*\|_1 \geq 0$, and since $\widehat{\boldsymbol{\Sigma}} \succeq \mathbf{0}$ and $\boldsymbol{\Omega}^* \succ \mathbf{0}$, it follows that $\langle \widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Omega}^*\rangle \geq 0$. Therefore, $\|\boldsymbol{\Omega}^*\|_2 \leq \frac{N}{\rho}$. □

Even though the choice of the linear operator in eq.(2.1) does not affect the positive definiteness properties of the estimated precision matrix or the optimization algorithm, in the following Section 2.4, we discuss positive definiteness properties and develop an optimization algorithm for the specific case of the discrete derivative operator $\mathbf{D}$.

2.4 Coordinate-Direction Descent Algorithm

Positive definiteness of the precision matrix is a necessary condition for the definition of a multivariate normal distribution. Furthermore, strict convexity is a very desirable property in optimization, since it ensures the existence of a unique global minimum. Notice that the penalized maximum likelihood estimation problem in eq.(2.1) is strictly convex due to the convexity properties of $\log\det\boldsymbol{\Omega}$ on the space of symmetric positive definite matrices [Boyd and Vandenberghe, 2006]. Maximization can be performed with respect to one row and column of the precision matrix $\boldsymbol{\Omega}$ at a time. Without loss of generality, we use the last row and column in our derivation, since permutation of rows and columns is always possible. Also, note that rows in $\mathbf{D}$ can be freely permuted without affecting the objective function. Let:

$$\boldsymbol{\Omega} = \left[\begin{array}{cc} \mathbf{W} & \mathbf{y} \\ \mathbf{y}^{\mathrm{T}} & z \end{array} \right] \quad , \quad \widehat{\boldsymbol{\Sigma}} = \left[\begin{array}{cc} \mathbf{S} & \mathbf{u} \\ \mathbf{u}^{\mathrm{T}} & v \end{array} \right] \quad , \quad \mathbf{D} = \left[\begin{array}{cc} \mathbf{D}_1 & \mathbf{0}_{M-L} \\ \mathbf{D}_2 & \mathbf{d}_3 \end{array} \right] \tag{2.6}$$

where $\mathbf{W}, \mathbf{S} \in \mathbb{R}^{N-1\times N-1}$, $\mathbf{y}, \mathbf{u} \in \mathbb{R}^{N-1}$, $\mathbf{d}_3 \in \mathbb{R}^{L}$ is a vector with all entries different than zero, which requires a permutation of rows in $\mathbf{D}$, $\mathbf{D}_1 \in \mathbb{R}^{M-L\times N-1}$ and $\mathbf{D}_2 \in \mathbb{R}^{L\times N-1}$.

In term of the variables $\mathbf{y}, z$ and the constant matrix $\mathbf{W}$, the penalized maximum likelihood estimation problem in eq.(2.1) can be reformulated as:

$$\max_{\boldsymbol{\Omega}\succ\mathbf{0}} \Big(\log(z - \mathbf{y}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{y}) - 2\mathbf{u}^{\mathrm{T}}\mathbf{y} - (v+\rho)z - 2\rho\|\mathbf{y}\|_1 - \tau\|\mathbf{A}\mathbf{y} - \mathbf{b}\|_1 \Big) \tag{2.7}$$

where $\|\mathbf{A}\mathbf{y} - \mathbf{b}\|_1$ can be written in an extended form:

$$\|\mathbf{A}\mathbf{y} - \mathbf{b}\|_1 = \|\mathbf{D}_1\mathbf{y}\|_1 + \|\mathbf{vec}(\mathbf{J}(\mathbf{D}_2) \circ (\mathbf{d}_3\mathbf{y}^{\mathrm{T}} + \mathbf{D}_2\mathbf{W}))\|_1 \tag{2.8}$$

Intuitively, the term $\|\mathbf{D}_1\mathbf{y}\|_1$ penalizes differences across different rows of $\boldsymbol{\Omega}$ which affect only values in $\mathbf{y}$, while the term $\|\mathbf{vec}(\mathbf{J}(\mathbf{D}_2) \circ (\mathbf{d}_3\mathbf{y}^{\mathrm{T}} + \mathbf{D}_2\mathbf{W}))\|_1$ penalizes differences across different columns of $\boldsymbol{\Omega}$ which affect values of $\mathbf{y}$ as well as $\mathbf{W}$.

It can be shown that the precision matrix $\boldsymbol{\Omega}$ is positive definite since its Schur complement $z - \mathbf{y}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{y}$ is positive. By maximizing eq.(2.7) with respect to z, we get:

$$z - \mathbf{y}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{y} = \frac{1}{v+\rho} \tag{2.9}$$

and since $v > 0$ and $\rho > 0$, this implies that the Schur complement in eq.(2.9) is positive.

Maximization with respect to one variable at a time leads to a strictly convex, non-smooth, piecewise quadratic function. By replacing the optimal value for z given by eq.(2.9) into the objective function in eq.(2.7), we get:

$$\min_{\mathbf{y}\in\mathbb{R}^{N-1}} \Big(\tfrac{1}{2}\mathbf{y}^{\mathrm{T}}(v+\rho)\mathbf{W}^{-1}\mathbf{y} + \mathbf{u}^{\mathrm{T}}\mathbf{y} + \rho\|\mathbf{y}\|_1 + \tfrac{\tau}{2}\|\mathbf{A}\mathbf{y} - \mathbf{b}\|_1 \Big) \tag{2.10}$$

Since the objective function in eq.(2.10) is non-smooth, its derivative is not continuous and therefore methods such as gradient descent cannot be applied. Although coordinate descent methods [Banerjee et al., 2006, Friedman et al., 2007b] are suitable when only sparseness is enforced, they are not when local constancy is encouraged. As shown in [Friedman et al., 2007a], when penalizing an ℓ_1-norm of differences, a coordinate descent algorithm can get stuck at sharp corners of the non-smooth optimization function; the resulting coordinates are stationary only under single-coordinate moves but not under diagonal moves involving two coordinates at a time.

For a discrete derivative operator $\mathbf{D}$ used in the penalized maximum likelihood estimation problem in eq.(2.1), it suffices to take directions involving either one variable $\mathbf{g} =$

$(0,\ldots,0,1,0,\ldots,0)^{\mathrm{T}}$ or two spatially neighboring variables $\mathbf{g} = (0,\ldots,0,1,0,\ldots,0,1,0,\ldots,0)^{\mathrm{T}}$ such that 1s appear in the position corresponding to the two neighbor variables. Finally, assuming an initial value $\mathbf{y}_0$ and a direction $\mathbf{g}$, the objective function in eq.(2.10) can be reduced to find t in $\mathbf{y}(t) = \mathbf{y}_0 + t\mathbf{g}$ such that it minimizes:

$$\begin{array}{l} \min_{t\in\mathbb{R}} \left(\frac{1}{2}pt^2 + qt + \sum_m r_m|t - s_m|\right) \\ p = (v+\rho)\mathbf{g}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{g} \quad , \quad q = ((v+\rho)\mathbf{W}^{-1}\mathbf{y}_0 + \mathbf{u})^{\mathrm{T}}\mathbf{g} \\ \mathbf{r} = \left[\begin{array}{c} \rho|\mathbf{g}| \\ \frac{\tau}{2}|\mathbf{A}\mathbf{g}| \end{array}\right] \quad , \quad \mathbf{s} = \left[\begin{array}{c} -\mathbf{Diag}(\mathbf{g})^{-1}(\mathbf{y}_0) \\ -\mathbf{Diag}(\mathbf{A}\mathbf{g})^{-1}(\mathbf{A}\mathbf{y}_0 - \mathbf{b}) \end{array}\right] \end{array} \tag{2.11}$$

For simplicity of notation, we assume that $\mathbf{r}, \mathbf{s} \in \mathbb{R}^M$ use only non-zero entries of $\mathbf{g}$ and $\mathbf{A}\mathbf{g}$ on its definition in eq.(2.11). We sort and remove duplicate values in $\mathbf{s}$, and propagate changes to $\mathbf{r}$ by adding the entries corresponding to the duplicate values in $\mathbf{s}$. Note that these apparent modifications do not change the objective function, but they simplify its optimization. The resulting minimization problem in eq.(2.11) is convex, non-smooth and piecewise quadratic. Furthermore, since the objective function is quadratic on each interval $[-\infty; s_1], [s_1; s_2], \ldots, [s_{M-1}; s_M], [s_M; +\infty]$, it admits a closed form solution. More formally, assume $s_0 = -\infty < s_1 < s_2 < \cdots < s_M < s_{M+1} = +\infty$, eq.(2.11) is equivalent to:

$$\begin{array}{l} \min_{t\in[s_{m-1};s_m],m} \left(\frac{1}{2}pt^2 + k_m t + l_m\right) \\ k_{m_1} = \left(q + \sum_{m_2<m_1} r_{m_2} - \sum_{m_2\geq m_1} r_{m_2}\right) \\ l_{m_1} = -\left(\sum_{m_2<m_1} r_{m_2}s_{m_2} - \sum_{m_2\geq m_1} r_{m_2}s_{m_2}\right) \end{array} \tag{2.12}$$

which has the closed form solution:

$$\begin{array}{l} t^* = \operatorname{argmin}_{t_m,m} \left(\frac{1}{2}pt_m^2 + k_m t_m + l_m\right) \\ t_m = \left\{\begin{array}{l} \frac{-k_m}{p}, \text{if } s_{m-1} \leq \frac{-k_m}{p} \leq s_m \\ s_{m-1}, \text{if } \frac{-k_m}{p} < s_{m-1} \\ s_m, \text{if } \frac{-k_m}{p} > s_m \end{array}\right. \end{array} \tag{2.13}$$

Algorithm 2.1 shows the coordinate-direction descent method in detail. A careful implementation of the algorithm allows obtaining a time complexity of $O(KN^3)$ for K iterations and N variables, in which $\mathbf{W}^{-1}$, $\mathbf{W}^{-1}\mathbf{y}$ and $\mathbf{A}\mathbf{y}$ are updated at each iteration. In our experiments, the algorithm converges quickly in usually $K = 10$ iterations. The polynomial dependency on the number of variables of $O(N^3)$ is expected since we cannot produce an algorithm faster than computing the inverse of the sample covariance in the case of an infinite sample.

Finally, in the spirit of [Banerjee et al., 2006], a method for reducing the size of the original problem is presented. Given a P-dimensional spatial neighborhood or manifold (e.g. $P = 1$ for silhouettes, $P = 2$ for a four-pixel neighborhood on 2D images, $P = 3$ for a six-pixel neighborhood on 3D images), the objective function in eq.(2.10) has the maximizer $\mathbf{y} = \mathbf{0}$ for variables on which $\|\mathbf{u}\|_\infty \leq \rho - P\tau$. Since this condition does not depend on specific entries in the iterative estimation of the precision matrix, this property can be used to reduce the size of the problem in advance by removing such variables.

Algorithm 2.1 Block Coordinate Descent for Local Constancy.

```
**Input:** $\widehat{\mathbf{\Sigma}} \succeq \mathbf{0}$, $\rho > 0$, $\tau > 0$, $\mathbf{D}$
Initialize $\mathbf{\Omega} = \mathbf{Diag}(\widehat{\mathbf{\Sigma}})^{-1}$
**for** each iteration $1, \dots, K$ and each variable $1, \dots, N$ **do**
  Split $\mathbf{\Omega}$ into $\mathbf{W}, \mathbf{y}, z$ and $\widehat{\mathbf{\Sigma}}$ into $\mathbf{S}, \mathbf{u}, v$ as described in eq.(2.6)
  Update $\mathbf{W}^{-1}$ by using the Sherman-Woodbury-Morrison formula (Note that when iterating from one variable to the next one, only one row and column change on matrix $\mathbf{W}$, see Appendix B)
  Transform local constancy regularization term from $\mathbf{D}$ into $\mathbf{A}$ and $\mathbf{b}$ as described in eq.(2.8)
  Compute $\mathbf{W}^{-1}\mathbf{y}$ and $\mathbf{A}\mathbf{y}$
  **for** each direction $\mathbf{g}$ involving either one variable or two spatially neighboring variables **do**
    Find $t$ that minimizes eq.(2.11) in closed form
    Update $\mathbf{y} \leftarrow \mathbf{y} + t\mathbf{g}$
    Update $\mathbf{W}^{-1}\mathbf{y} \leftarrow \mathbf{W}^{-1}\mathbf{y} + t\mathbf{W}^{-1}\mathbf{g}$
    Update $\mathbf{A}\mathbf{y} \leftarrow \mathbf{A}\mathbf{y} + t\mathbf{A}\mathbf{g}$
  **end for**
  Update $z \leftarrow \frac{1}{v+\rho} + \mathbf{y}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{y}$
**end for**
**Output:** $\mathbf{\Omega} \succ \mathbf{0}$
```

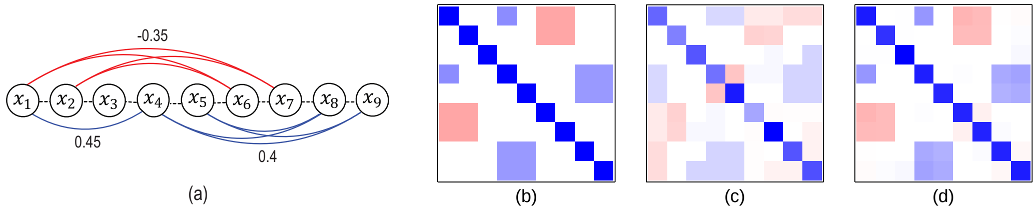


Figure 2.1: Synthetic model with locally constant interactions and learnt structures. (a) Ground truth model on an open contour manifold. Spatial neighbors are connected with black dashed lines. Positive interactions are shown in blue, negative interactions in red. The model contains two locally constant interactions between (x_1, x_2) and (x_6, x_7), and between (x_4, x_5) and (x_8, x_9), a not locally constant interaction between x_1 and x_4, and an independent variable x_3; (b) colored precision matrix of the ground truth, red for negative entries, blue for positive entries; learnt structure from (c) small and (d) large datasets. Note that for large datasets all connections are correctly recovered.

2.5 Experimental Results

We begin with a small synthetic example to test the ability of the method for recovering the ground truth structure from data, in a complex scenario in which our method has to deal with both locally and not locally constant interactions as well as independent variables. The ground truth Gaussian graphical model is shown in Figure 2.1 and it contains 9 variables arranged in an open contour manifold.

In order to measure the closeness of the recovered models to the ground truth, we measure the Kullback-Leibler divergence, average precision (one minus the fraction of falsely included edges), average recall (one minus the fraction of falsely excluded edges) as well as the Frobenius norm between the recovered model and the ground truth. For comparison purposes, we picked two of the state-of-the-art structure learning techniques: covariance selection [Banerjee et al., 2006] and graphical lasso [Friedman et al., 2007b], since it has been shown theoretically and experimentally that they both converge to the maximum like-

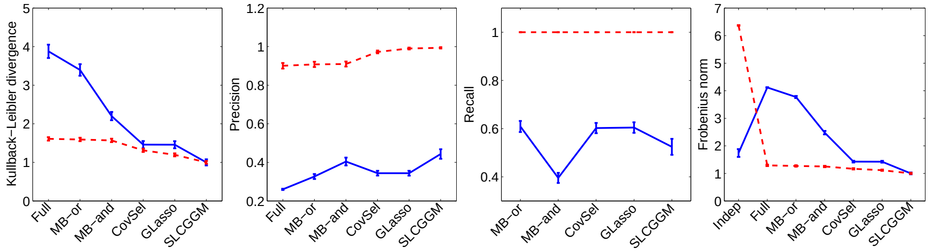


Figure 2.2: Kullback-Leibler divergence with respect to the best method, average precision, recall and Frobenius norm between the recovered model and the ground truth. Our method (SLCGGM) outperforms the fully connected model (Full), Meinshausen-Bühlmann approximation (MB-or, MB-and), covariance selection (CovSel), graphical lasso (GLasso) for small datasets (in blue solid line) and for large datasets (in red dashed line). The fully independent model (Indep) resulted in relative divergences of 2.49 for small and 113.84 for large datasets.

lihood estimator. We also test the Meinshausen-Bühlmann approximation [Meinshausen and Bühlmann, 2006]. The fully connected as well as fully independent model are also included as baseline methods.

Two different scenarios are tested: small datasets of four samples, and large datasets of 400 samples. Under each scenario, 50 datasets are randomly generated from the ground truth Gaussian graphical model. It can be concluded from Figure 2.2 that our method outperforms the state-of-the-art structure learning techniques both for small and large datasets. This is due to the fact that the ground truth data contains locally constant interactions, and our method imposes a prior for local constancy. Although this is a complex scenario which also contains not locally constant interactions as well as an independent variable, our method can recover a more plausible model when compared to other methods. Note that even though other methods may exhibit a higher recall for small datasets, our method consistently recovers a better probability distribution.

A visual comparison of the ground truth versus the best recovered model by our method from small and large datasets is shown in Figure 2.1. The image shows the precision matrix in which red squares represent negative entries, while blue squares represent positive entries. There is very little difference between the ground truth and the recovered model from large datasets. Although the model is not fully recovered from small datasets, our technique performs better than the Meinshausen-Bühlmann approximation, covariance selection and graphical lasso in Figure 2.2.

In the following experiments, we demonstrate the ability of our method to discover useful structures from real-world datasets. Datasets with a diverse nature of probabilistic relationships are included in our experiments: cardiac MRI, a walking sequence and functional brain MRI. We used short-axis *cardiac MRI* collected by Deux et al. [2008]. The segmentation of the myocarde at the end diastole performed by an expert was used for cropping. Preprocessing of the dataset was performed in DROP (`http://campar.in.tum.de/Main/Drop`) for computing the displacements of the pixels from the initial reference frame. The cardiac MRI sequence contains 24 images of 100 × 100 pixels. After preprocessing, the number of pixels that correspond to the heart was reduced to 122. We applied our algorithm with sparseness parameter $\rho = 0.35$, local constancy parameter $\tau = 0.05$ and $K = 10$ iterations. We also used a *walking sequence* from the Human Identification at a Distance dataset

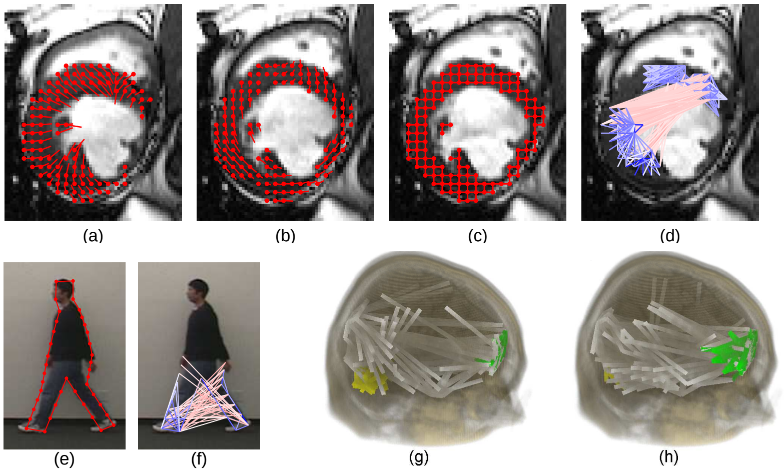


Figure 2.3: Results on real-world datasets: cardiac MRI displacement (a) at full contraction and (b) at full expansion, (c) 2D spatial manifold and (d) learnt structure, which captures contraction and expansion (in red), and similar displacements between neighbor pixels (in blue); (e) silhouette manifold and (f) learnt structure from a manually labeled walking sequence, showing similar displacements from each independent leg (in blue) and opposite displacements between both legs as well as between hands and feet (in red); and structures learnt from functional brain MRI in a monetary reward task for (g) drug addicted subjects with more connections in the cerebellum (in yellow) versus (h) control subjects with more connections in the prefrontal cortex (in green).

(`http://www.cc.gatech.edu/cpl/projects/hid/`). The silhouette consisting of 40 landmarks was manually labeled in a video sequence of 79 frames. We applied our algorithm with sparseness parameter $\rho = 2.5$, local constancy parameter $\tau = 0.25$ and $K = 10$ iterations. Finally, we used *functional brain MRI* collected by Goldstein et al. [2007]. The time series consists of 87 frames taken every 3.5 seconds. The dataset contains 28 subjects: 16 drug-addicted and 12 healthy non-drug-using control individuals. Preprocessing of the dataset was performed in SPM2 (`http://www.fil.ion.ucl.ac.uk/spm/`), and it included deforming all time series to the same spatial reference template (Talairach space), spatial smoothing, cropping and regular sampling. Each subject has a sequence of 87 images of $53 \times 63 \times 46$ voxels. After preprocessing, the number of voxels was reduced to 869. We applied our algorithm with sparseness parameter $\rho = 0.15$, local constancy parameter $\tau = 0.01$ and $K = 10$ iterations.

From the cardiac MRI [Deux et al., 2008], our method recovers global deformation in the form of rotation and shrinking; from the walking sequence, our method finds the long range interactions between different parts; and from the functional brain MRI, our method recovers functional interactions between different regions and discover differences in processing monetary rewards between cocaine addicted subjects versus healthy control subjects. Each dataset is also diverse in the type of spatial neighborhood: one-dimensional for silhouettes in a walking sequence, two-dimensional for cardiac MRI and three-dimensional for functional brain MRI.

Cross-validation was performed in order to measure the generalization performance of our method in estimating the underlying distribution. Each dataset was randomly split into five sets. On each round, four sets were used for training and the remaining set was used for

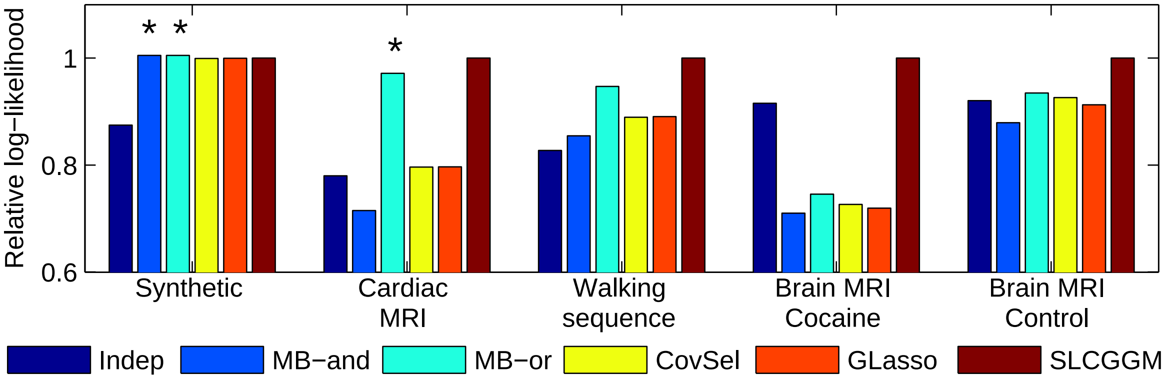


Figure 2.4: Cross-validated log-likelihood on the testing set. Our method (SLCGGM) outperforms the Meinshausen-Bühlmann approximation (MB-and, MB-or), covariance selection (CovSel), graphical lasso (GLasso) and the fully independent model (Indep). Bars marked with an asterisk are not statistically significantly different from our method.

measuring the log-likelihood. We tested for statistical significance by using the likelihood ratio test as follows: given the log-likelihood L_1 of a model with κ_1 parameters (number of non-zero entries in the precision matrix), and the log-likelihood L_0 of a simpler model with $\kappa_0 < \kappa_1$ parameters, at significance level $1 - \alpha = 0.95$ we reject the simpler model if $2(L_1 - L_0) \geq \chi^2_{(\alpha=0.05, DOF=\kappa_1-\kappa_0)}$.

Figure 2.4 shows that our method consistently outperforms techniques that encourage sparsity only. This is strong evidence that datasets that are measured over a spatial manifold are locally constant, as well as that our method is a good regularization technique that avoids over-fitting and allows for better generalization. Another interesting fact is that for the brain MRI dataset, which is high dimensional and contains a small number of samples, the model that assumes full independence performed better than the Meinshausen-Bühlmann approximation, covariance selection and graphical lasso. Similar observations has been already made in [Domingos and Pazzani, 1997, Friedman et al., 1997] where it was found that assuming independence often performs better than learning dependencies among variables.

2.6 Concluding Remarks

In this chapter, we proposed local constancy for Gaussian graphical models, which encourages finding probabilistic connectivities between two close or distant clusters of variables, instead of between isolated variables. We introduced an ℓ_1-norm penalty for local constancy into a strictly convex maximum likelihood estimation. Furthermore, we proposed an efficient optimization algorithm and proved that our method guarantees positive definiteness of the estimated precision matrix. We tested the ability of our method to recover the ground truth structure from data, in a complex scenario with locally and not locally constant interactions as well as independent variables. We also tested the generalization performance of our method in a wide range of complex real-world datasets with a diverse nature of probabilistic relationships as well as neighborhood type.

Chapter 3

Learning Independent Gaussian MRFs: Variable Selection

In the previous chapter, we proposed a prior for learning structures which is suitable for spatial datasets. In this chapter, we propose more general priors, not only for spatial datasets, but for any dataset with a large number of variables.

We present a *variable-selection structure learning* approach for Gaussian graphical models. Unlike standard sparseness promoting techniques, our method aims at selecting the most-important variables besides simply sparsifying the set of edges. Through simulations, we show that our method outperforms the state-of-the-art in recovering the ground truth model. Our method also exhibits better generalization performance in a wide range of complex real-world datasets: brain fMRI, gene expression, NASDAQ stock prices and world weather. We also show that our resulting networks are more interpretable in the context of brain fMRI analysis, while retaining discriminability. From an optimization perspective, we show that a block coordinate descent method generates a sequence of positive definite solutions. Thus, we reduce the original problem into a sequence of strictly convex (ℓ_1,ℓ_p) regularized quadratic minimization subproblems for $p \in \{2, \infty\}$. Our algorithm is well founded since the optimal solution of the maximization problem is unique and bounded.

3.1 Introduction

In this chapter, we enforce a particular form of sparseness: that only a small number of nodes in the graphical model interact with each other. Intuitively, we want to select these "important" nodes. However, methods for sparsifying network structure [Banerjee et al., 2006, Friedman et al., 2007b, Meinshausen and Bühlmann, 2006, Yuan and Lin, 2007] do not directly promote variable selection, i.e. group-wise elimination of all edges adjacent to an "unimportant" node. Variable selection in graphical models present several advantages. From a computational point of view, reducing the number of variables can significantly reduce the number of precision-matrix parameters. Moreover, group-wise edge elimination may serve as a more aggressive regularization, removing all "noisy" edges associated with nuisance variables at once, and potentially leading to better generalization performance, especially if, indeed, the underlying problem structure involves only a limited number of "important" variables. Finally, variable selection improves interpretability of the graphical model: for example, when learning a graphical model of brain area connectivity, variable selection may help to localize brain areas most relevant to particular mental states.

It has been demonstrated that Gaussian graphical models can already achieve promising predictive performance on mental-state prediction tasks [Cecchi et al., 2009] using a small number of variables (voxels) pre-selected via a simple univariate ranking of each variable's relevance to the response (class label). Going beyond such univariate ranking, and embedding variable-selection process into model-building guided by likelihood maximization, can result into even more accurate models and identify even more informative subsets of variables (i.e. brain regions).

Our contribution is to develop variable-selection in the context of learning sparse Gaussian graphical models. To achieve this, we add an $\ell_{1,p}$-norm regularization term to the maximum likelihood estimation problem, for $p \in \{2, \infty\}$. We optimize this problem through a block coordinate descent method which yields sparse and positive definite estimates. We show that our method outperforms the state-of-the-art in recovering the ground truth model through synthetic experiments. We also show that our structures have higher test log-likelihood than competing methods, in a wide range of complex real-world datasets: brain fMRI, gene expression, NASDAQ stock prices and world weather. In particular, in the context of brain fMRI analysis, we show that our method produces more interpretable models that involve few brain areas, unlike standard sparseness promoting techniques which produce hard-to-interpret networks involving most of the brain. Moreover, our structures are as good as standard sparseness promoting techniques, when used for classification purposes.

Section 3.2 introduces techniques for learning Gaussian graphical models from data. Section 3.3 sets up the $\ell_{1,p}$-regularized maximum likelihood problem and discusses its properties. Section 3.4 describes our block coordinate descent method. Experimental results are in Section 3.5.

3.2 Background

In Section 1.4, we introduced Gaussian graphical models as well as techniques for learning sparse Gaussian graphical models through ℓ_1 regularization, such as: *covariance selection* [Banerjee et al., 2006], *graphical lasso* [Friedman et al., 2007b] and the *Meinshausen-Bühlmann approximation* [Meinshausen and Bühlmann, 2006].

Besides sparseness, several regularizers have been proposed for Gaussian graphical models, for enforcing diagonal structure [Levina et al., 2008], spatial coherence [Honorio et al., 2009], common structure among multiple tasks [Honorio and Samaras, 2010], or sparse changes in controlled experiments [Zhang and Wang, 2010]. In particular, different group sparse priors have been proposed for enforcing block structure for known block-variable assignments [Duchi et al., 2008a, Schmidt et al., 2009] and unknown block-variable assignments [Marlin and K.Murphy, 2009, Marlin et al., 2009], or power law regularization in scale free networks [Liu and Ihler, 2011].

Variable selection has been applied to very diverse problems, such as linear regression [Tibshirani, 1996], classification [Chan et al., 2007, Lee et al., 2006b, Duchi and Singer, 2009a] and reinforcement learning [Parr et al., 2008].

3.3 Preliminaries

In this section, we set up the problem and discuss some of its properties.

3.3.1 Problem Setup

We propose priors that are motivated from the variable selection literature from regression and classification, such as group lasso [Yuan and Lin, 2006, Meier et al., 2008, Obozinski et al., 2010] which imposes an $\ell_{1,2}$-norm penalty, and simultaneous lasso [Turlach et al., 2005, Tropp, 2006] which imposes an $\ell_{1,\infty}$-norm penalty.

Recall that an edge in a Gaussian graphical model corresponds to a non-zero entry in the precision matrix. We promote variable selection by learning a structure with a small number of nodes that interact with each other, or equivalently a large number of nodes that are disconnected from the rest of the graph. For each disconnected node, its corresponding row in the precision matrix (or column given that it is symmetric) contains only zeros (except for the diagonal). Therefore, the use of row-level regularizers such as the $\ell_{1,p}$-norm are natural in our context. Note that our goal differs from sparse Gaussian graphical models, in which sparseness is imposed at the edge level only. We additionally impose sparseness at the node level, which promotes conditional independence of variables with respect to all other variables.

Given a dense sample covariance matrix $\widehat{\mathbf{\Sigma}} \succeq \mathbf{0}$, we learn a precision matrix $\mathbf{\Omega} \in \mathbb{R}^{N\times N}$ for N variables. The *variable-selection structure learning problem* is defined as:

$$\max_{\mathbf{\Omega}\succ\mathbf{0}} \left(\log\det\mathbf{\Omega} - \langle\widehat{\mathbf{\Sigma}},\mathbf{\Omega}\rangle - \rho\|\mathbf{\Omega}\|_1 - \tau\|\mathbf{\Omega}\|_{1,p}\right) \tag{3.1}$$

for $\rho > 0$, $\tau > 0$ and $p \in \{2,\infty\}$. The term $\log\det\mathbf{\Omega} - \langle\widehat{\mathbf{\Sigma}},\mathbf{\Omega}\rangle$ is the Gaussian log-likelihood. $\|\mathbf{\Omega}\|_1$ encourages sparseness of the precision matrix or conditional independence among variables. The last term $\|\mathbf{\Omega}\|_{1,p}$ is our variable selection regularizer, and it is defined as:

$$\|\mathbf{\Omega}\|_{1,p} = \sum_n \|(\omega_{n,1},\ldots,\omega_{n,n-1},\omega_{n,n+1},\ldots,\omega_{n,N})\|_p \tag{3.2}$$

In a technical report, Friedman et al. [2010] proposed an optimization problem that is similar to eq.(3.1). The main differences are that their model does not promote sparseness, and that they do not solve the original maximum likelihood problem, but instead build upon an approximation (pseudo-likelihood) approach of Meinshausen and Bühlmann [2006] based on independent linear regression problems. Finally, note that regression based methods such as [Meinshausen and Bühlmann, 2006] have been already shown in [Friedman et al., 2007b] to have worse performance than solving the original maximum likelihood problem. In this chapter, we solve the original maximum likelihood problem.

3.3.2 Bounds

In what follows, we discuss uniqueness and boundedness of the optimal solution of our problem.

Lemma 3.1. *For $\rho > 0$, $\tau > 0$, the variable-selection structure learning problem in eq.*(3.1) *is a maximization problem with concave (but not strictly concave) objective function and convex constraints.*

Proof. The Gaussian log-likelihood is concave, since $\log\det$ is concave on the space of symmetric positive definite matrices, and since the linear operator $\langle\cdot,\cdot\rangle$ is also concave. Both regularization terms, the negative ℓ_1-norm as well as the negative $\ell_{1,p}$-norm defined in eq.(3.2) are non-smooth concave functions. Finally, $\mathbf{\Omega} \succ \mathbf{0}$ is a convex constraint. □

For clarity of exposition, we assume that the diagonals of $\mathbf{\Omega}$ are penalized by our variable selection regularizer defined in eq.(3.2).

Theorem 3.2. *For $\rho > 0$, $\tau > 0$, the optimal solution to the variable-selection structure learning problem in eq.*(3.1) *is unique and bounded as follows:*

$$\left(\frac{1}{\|\widehat{\mathbf{\Sigma}}\|_2 + N\rho + N^{1/p'}\tau}\right)\mathbf{I} \preceq \mathbf{\Omega}^* \preceq \left(\frac{N}{\max(\rho,\tau)}\right)\mathbf{I} \tag{3.3}$$

where $\ell_{p'}$-norm is the dual of the ℓ_p-norm, i.e. $(p = 2, p' = 2)$ or $(p = \infty, p' = 1)$.

Proof. By using the identity for dual norms $\kappa\|\mathbf{c}\|_p = \max_{\|\mathbf{d}\|_{p'}\leq\kappa} \mathbf{d}^{\mathrm{T}}\mathbf{c}$ in eq.(3.1), we get:

$$\max_{\mathbf{\Omega}\succ\mathbf{0}} \min_{\substack{\|\mathbf{A}\|_\infty\leq\rho \\ \|\mathbf{B}\|_{\infty,p'}\leq\tau}} \left(\log\det\mathbf{\Omega} - \langle\widehat{\mathbf{\Sigma}} + \mathbf{A} + \mathbf{B}, \mathbf{\Omega}\rangle\right) \tag{3.4}$$

where $\|B\|_{\infty,p'} = \max_n \|(b_{n,1},\ldots,b_{n,N})\|_{p'}$. By virtue of Sion's minimax theorem, we can swap the order of max and min. Furthermore, note that the optimal solution of the inner equation is given by $\mathbf{\Omega} = (\widehat{\mathbf{\Sigma}} + \mathbf{A} + \mathbf{B})^{-1}$. By replacing this solution in eq.(3.4), we get the dual problem of eq.(3.1):

$$\min_{\substack{\|\mathbf{A}\|_\infty\leq\rho \\ \|\mathbf{B}\|_{\infty,p'}\leq\tau}} \left(-\log\det(\widehat{\mathbf{\Sigma}} + \mathbf{A} + \mathbf{B}) - N\right) \tag{3.5}$$

In order to find a lower bound for the minimum eigenvalue of $\mathbf{\Omega}^*$, note that $\|\mathbf{\Omega}^{*-1}\|_2 = \|\widehat{\mathbf{\Sigma}}+\mathbf{A}+\mathbf{B}\|_2 \leq \|\widehat{\mathbf{\Sigma}}\|_2+\|\mathbf{A}\|_2+\|\mathbf{B}\|_2 \leq \|\widehat{\mathbf{\Sigma}}\|_2+N\|\mathbf{A}\|_\infty+N^{1/p'}\|\mathbf{B}\|_{\infty,p'} \leq \|\widehat{\mathbf{\Sigma}}\|_2+N\rho+N^{1/p'}\tau$. (Here we used $\|\mathbf{B}\|_2 \leq N^{1/p'}\|\mathbf{B}\|_{\infty,p'}$ as shown in Appendix C)

In order to find an upper bound for the maximum eigenvalue of $\mathbf{\Omega}^*$, note that, at optimum, the primal-dual gap is zero:

$$-N + \langle\widehat{\mathbf{\Sigma}}, \mathbf{\Omega}^*\rangle + \rho\|\mathbf{\Omega}^*\|_1 + \tau\|\mathbf{\Omega}^*\|_{1,p} = 0 \tag{3.6}$$

The upper bound is found as follows: $\|\mathbf{\Omega}^*\|_2 \leq \|\mathbf{\Omega}^*\|_{\mathfrak{F}} \leq \|\mathbf{\Omega}^*\|_1 = (N - \langle\widehat{\mathbf{\Sigma}}, \mathbf{\Omega}^*\rangle - \tau\|\mathbf{\Omega}^*\|_{1,p})/\rho$. Note that $\tau\|\mathbf{\Omega}^*\|_{1,p} \geq 0$, and since $\widehat{\mathbf{\Sigma}} \succeq \mathbf{0}$ and $\mathbf{\Omega}^* \succ \mathbf{0}$, it follows that $\langle\widehat{\mathbf{\Sigma}}, \mathbf{\Omega}^*\rangle \geq 0$. Therefore, $\|\mathbf{\Omega}^*\|_2 \leq \frac{N}{\rho}$. In a similar fashion, $\|\mathbf{\Omega}^*\|_2 \leq \|\mathbf{\Omega}^*\|_{1,p} = (N-\langle\widehat{\mathbf{\Sigma}}, \mathbf{\Omega}^*\rangle-\rho\|\mathbf{\Omega}^*\|_1)/\tau$. (Here we used $\|\mathbf{\Omega}^*\|_2 \leq \|\mathbf{\Omega}^*\|_{1,p}$ as shown in Appendix C). Note that $\rho\|\mathbf{\Omega}^*\|_1 \geq 0$ and $\langle\widehat{\mathbf{\Sigma}}, \mathbf{\Omega}^*\rangle \geq 0$. Therefore, $\|\mathbf{\Omega}^*\|_2 \leq \frac{N}{\tau}$. □

3.4 Block Coordinate Descent Method

Since the objective function in eq.(3.1) contains a non-smooth regularizer, methods such as gradient descent cannot be applied. On the other hand, subgradient descent methods very rarely converge to non-smooth points [Duchi and Singer, 2009b]. In our problem, these non-smooth points correspond to zeros in the precision matrix, are often the true minima of the objective function, and are very desirable in the solution because they convey information of conditional independence among variables.

We apply block coordinate descent method on the primal problem [Honorio et al., 2009, Honorio and Samaras, 2010], unlike covariance selection [Banerjee et al., 2006] and graphical lasso [Friedman et al., 2007b] which optimize the dual. Optimization of the dual problem in eq.(3.5) by a block coordinate descent method can be done with quadratic programming for $p = \infty$ but not for $p = 2$ (i.e. the objective function is quadratic for $p \in \{2, \infty\}$, the constraints are linear for $p = \infty$ and quadratic for $p = 2$). Optimization of the primal problem provides the same efficient framework for $p \in \{2, \infty\}$. We point out that a projected subgradient method as in [Duchi et al., 2008a] cannot be applied since our regularizer does not decompose into disjoint subsets. Our problem contains a positive definiteness constraint and therefore it does not fall in the general framework of [Yuan and Lin, 2006, Meier et al., 2008, Obozinski et al., 2010, Quattoni et al., 2009, Turlach et al., 2005, Tropp, 2006] which consider unconstrained problems only. Finally, more recent work of [Chen et al., 2011, Mairal et al., 2010] consider subsets with overlap, but it does still consider unconstrained problems only.

Theorem 3.3. *The block coordinate descent method for the variable-selection structure learning problem in eq.*(3.1) *generates a sequence of positive definite solutions.*

Proof. Maximization can be performed with respect to one row and column of all precision matrices $\mathbf{\Omega}$ at a time. Without loss of generality, we use the last row and column in our derivation. Let:

$$\mathbf{\Omega} = \begin{bmatrix} \mathbf{W} & \mathbf{y} \\ \mathbf{y}^{\mathrm{T}} & z \end{bmatrix} \quad , \quad \widehat{\mathbf{\Sigma}} = \begin{bmatrix} \mathbf{S} & \mathbf{u} \\ \mathbf{u}^{\mathrm{T}} & v \end{bmatrix} \tag{3.7}$$

where $\mathbf{W}, \mathbf{S} \in \mathbb{R}^{N-1 \times N-1}$, $\mathbf{y}, \mathbf{u} \in \mathbb{R}^{N-1}$.

In terms of the variables $\mathbf{y}, z$ and the constant matrix $\mathbf{W}$, the variable-selection structure learning problem in eq.(3.1) can be reformulated as:

$$\max_{\mathbf{\Omega} \succ \mathbf{0}} \begin{pmatrix} \log(z - \mathbf{y}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{y}) - 2\mathbf{u}^{\mathrm{T}}\mathbf{y} - (v + \rho)z \\ -2\rho\|\mathbf{y}\|_1 - \tau\|\mathbf{y}\|_p - \tau \sum_n \|(y_n, t_n)\|_p \end{pmatrix} \tag{3.8}$$

where $t_n = \|(w_{n,1}, \ldots, w_{n,n-1}, w_{n,n+1}, \ldots, w_{n,N})\|_p$.

If $\mathbf{\Omega}$ is a symmetric matrix, according to the Haynsworth inertia formula, $\mathbf{\Omega} \succ \mathbf{0}$ if and only if its Schur complement $z - \mathbf{y}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{y} > 0$ and $\mathbf{W} \succ \mathbf{0}$. By maximizing eq.(3.8) with respect to z, we get:

$$z - \mathbf{y}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{y} = \frac{1}{v + \rho} \tag{3.9}$$

and since $v > 0$ and $\rho > 0$, this implies that the Schur complement in eq.(3.9) is positive. Finally, in our iterative optimization, it suffices to initialize $\mathbf{\Omega}$ to a matrix known to be positive definite, e.g. a diagonal matrix with positive elements. □

Theorem 3.4. *The block coordinate descent method for the variable-selection structure learning problem in eq.*(3.1) *is equivalent to solving a sequence of strictly convex* $(\ell_1,\ell_{1,p})$ *regularized quadratic subproblems for* $p \in \{2,\infty\}$*:*

$$\min_{\mathbf{y}\in\mathbb{R}^{N-1}} \begin{pmatrix} \frac{1}{2}\mathbf{y}^{\mathrm{T}}(v+\rho)\mathbf{W}^{-1}\mathbf{y} + \mathbf{u}^{\mathrm{T}}\mathbf{y} \\ +\rho\|\mathbf{y}\|_1 + \frac{\tau}{2}\|\mathbf{y}\|_p + \frac{\tau}{2}\sum_n \|(y_n,t_n)\|_p \end{pmatrix} \tag{3.10}$$

Proof. By replacing the optimal z given by eq.(3.9) into the objective function in eq.(3.8), we get eq.(3.10). Since $\mathbf{W} \succ \mathbf{0} \Rightarrow \mathbf{W}^{-1} \succ \mathbf{0}$, hence eq.(3.10) is strictly convex. □

Lemma 3.5. *If* $\|\mathbf{u}\|_\infty \leq \rho + \tau/(2(N-1)^{1/p'})$ *or* $\|\mathbf{u}\|_{p'} \leq \rho + \tau/2$*, the* $(\ell_1,\ell_{1,p})$ *regularized quadratic problem in eq.*(3.10) *has the minimizer* $\mathbf{y}^* = \mathbf{0}$*.*

Proof. Note that since $\mathbf{W} \succ \mathbf{0} \Rightarrow \mathbf{W}^{-1} \succ \mathbf{0}$, $\mathbf{y}^* = \mathbf{0}$ is the minimizer of the quadratic part of eq.(3.10). It suffices to prove that the remaining part is also minimized for $\mathbf{y}^* = \mathbf{0}$, i.e. $\mathbf{u}^{\mathrm{T}}\mathbf{y} + \rho\|\mathbf{y}\|_1 + \frac{\tau}{2}\|\mathbf{y}\|_p + \frac{\tau}{2}\sum_n \|(y_n,t_n)\|_p \geq \frac{\tau}{2}\sum_n t_n$ for an arbitrary $\mathbf{y}$. The lower bound comes from setting $\mathbf{y}^* = \mathbf{0}$ in eq.(3.10) and by noting that $(\forall n)\ t_n > 0$.

By using lower bounds $\sum_n \|(y_n,t_n)\|_p \geq \sum_n t_n$ and either $\|\mathbf{y}\|_p \geq \|\mathbf{y}\|_1/(N-1)^{1/p'}$ or $\|\mathbf{y}\|_1 \geq \|\mathbf{y}\|_p$, we modify the original claim into a stronger one, i.e. $\mathbf{u}^{\mathrm{T}}\mathbf{y} + (\rho + \tau/(2(N-1)^{1/p'}))\|\mathbf{y}\|_1 \geq 0$ or $\mathbf{u}^{\mathrm{T}}\mathbf{y} + (\rho+\tau/2)\|y\|_p \geq 0$. Finally, by using the identity for dual norms $\kappa\|\mathbf{y}\|_p = \max_{\|\mathbf{d}\|_{p'}\leq\kappa} \mathbf{d}^{\mathrm{T}}\mathbf{y}$, we have the condition $\max_{\|\mathbf{d}\|_\infty \leq \rho+\tau/(2(N-1)^{1/p'})} (\mathbf{u}+\mathbf{d})^{\mathrm{T}}\mathbf{y} \geq 0$ or the condition $\max_{\|\mathbf{d}\|_{p'}\leq\rho+\tau/2} (\mathbf{u}+\mathbf{d})^{\mathrm{T}}\mathbf{y} \geq 0$, which proves our claim. □

Remark 3.6. *By using Lemma 3.5, we can reduce the size of the original problem by removing variables in which this condition holds, since it only depends on the dense sample covariance matrix.*

Theorem 3.7. *The coordinate descent method for the* $(\ell_1,\ell_{1,p})$ *regularized quadratic problem in eq.*(3.10) *is equivalent to solving a sequence of strictly convex* (ℓ_1,ℓ_p) *regularized quadratic subproblems:*

$$\min_x \left(\frac{1}{2}qx^2 - cx + \rho|x| + \frac{\tau}{2}\|(x,a)\|_p + \frac{\tau}{2}\|(x,b)\|_p\right) \tag{3.11}$$

Proof. Without loss of generality, we use the last row and column in our derivation, since permutation of rows and columns is always possible. Let:

$$\mathbf{W}^{-1} = \begin{bmatrix} \mathbf{H}_{11} & \mathbf{h}_{12} \\ \mathbf{h}_{12}{}^{\mathrm{T}} & h_{22} \end{bmatrix} \quad , \quad \mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ x \end{bmatrix} \quad , \quad \mathbf{u} = \begin{bmatrix} \mathbf{u}_1 \\ u_2 \end{bmatrix} \tag{3.12}$$

where $\mathbf{H}_{11} \in \mathbb{R}^{N-2\times N-2}$, $\mathbf{h}_{12}, \mathbf{y}_1, \mathbf{u}_1 \in \mathbb{R}^{N-2}$.

In terms of the variable x and the constants $q = (v+\rho)h_{22}$, $c = -((v+\rho)\mathbf{h}_{12}{}^{\mathrm{T}}\mathbf{y}_1 + u_2)$, $a = \|\mathbf{y}_1\|_p$, $b = t_n$, the $(\ell_1,\ell_{1,p})$ regularized quadratic problem in eq.(3.10) can be reformulated as in eq.(3.11). Moreover, since $v > 0 \wedge \rho > 0 \wedge h_{22} > 0 \Rightarrow q > 0$, and therefore eq.(3.11) is strictly convex. □

For $p = \infty$, eq.(3.11) has five points in which the objective function is non-smooth, i.e. $x \in \{-\max(a,b), -\min(a,b), 0, \min(a,b), \max(a,b)\}$. Furthermore, since the objective function is quadratic on each interval, it admits a closed form solution.

For $p = 2$, eq.(3.11) has only one non-smooth point, i.e. $x = 0$. Given the objective function $f(x)$, we first compute the left derivative $\partial_- f(0) = -c - \rho$ and the right derivative $\partial_+ f(0) = -c + \rho$. If $\partial_- f(0) \leq 0 \wedge \partial_+ f(0) \geq 0 \Rightarrow x^* = 0$. If $\partial_- f(0) > 0 \Rightarrow x^* < 0$ and we use the one-dimensional *Newton-Raphson method* for finding x^*. If $\partial_+ f(0) < 0 \Rightarrow x^* > 0$. For numerical stability, we add a small $\varepsilon > 0$ to the ℓ_2-norms by using $\sqrt{x^2 + a^2 + \varepsilon}$ instead of $\|(x, a)\|_2$.

Algorithm 3.1 shows the block coordinate descent method in detail. A careful implementation leads to a time complexity of $\mathcal{O}(KN^3)$ for K iterations and N variables. In our experiments, the algorithm converges quickly in usually $K = 10$ iterations. Polynomial dependence $\mathcal{O}(N^3)$ on the number of variables is expected since no algorithm can be faster than computing the inverse of the sample covariance in the case of an infinite sample.

Algorithm 3.1 Block Coordinate Descent for Variable Selection.

```
Input: $\widehat{\mathbf{\Sigma}} \succeq \mathbf{0}$, $\rho > 0$, $\tau > 0$, $p \in \{2, \infty\}$
Initialize $\mathbf{\Omega} = \mathbf{Diag}(\widehat{\mathbf{\Sigma}})^{-1}$
for each iteration $1, \dots, K$ and each variable $1, \dots, N$ do
  Split $\mathbf{\Omega}$ into $\mathbf{W}, \mathbf{y}, z$ and $\widehat{\mathbf{\Sigma}}$ into $\mathbf{S}, \mathbf{u}, v$ as described in eq.(3.7)
  Update $\mathbf{W}^{-1}$ by using the Sherman-Woodbury-Morrison formula (Note that when iterating from one variable to the next one, only one row and column change on matrix $\mathbf{W}$, see Appendix B)
  for each variable $1, \dots, N-1$ do
    Split $\mathbf{W}^{-1}, \mathbf{y}, \mathbf{u}$ as in eq.(3.12)
    Solve the $(\ell_1, \ell_p)$ regularized quadratic problem in closed form ($p = \infty$) or by using the Newton-Raphson method ($p = 2$)
  end for
  Update $z \leftarrow \frac{1}{v+\rho} + \mathbf{y}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{y}$
end for
Output: $\mathbf{\Omega} \succ \mathbf{0}$
```

3.5 Experimental Results

We test with a synthetic example the ability of the method to recover ground truth structure from data. The model contains $N \in \{50, 100, 200\}$ variables. For each of 50 repetitions, we first select a proportion of “connected” nodes (either 0.2,0.5,0.8) from the N variables. The unselected (i.e. “disconnected”) nodes do not participate in any edge of the ground truth model. We then generate edges among the connected nodes with a required density (either 0.2,0.5,0.8), where each edge weight is generated uniformly at random from $\{-1, +1\}$. We ensure positive definiteness of $\mathbf{\Omega}_g$ by verifying that its minimum eigenvalue is at least 0.1. We then generate a dataset of 50 samples. We model the ratio $\bar{\sigma}_c/\bar{\sigma}_d$ between the standard deviation of connected versus disconnected nodes. In the “high variance confounders” regime, $\bar{\sigma}_c/\bar{\sigma}_d = 1$ which means that on average connected and disconnected variables have the same standard deviation. In the “low variance confounders” regime, $\bar{\sigma}_c/\bar{\sigma}_d = 10$ which means that on average the standard deviation of a connected variable is 10 times the one of a disconnected variable. Variables with low variance produce higher values in the precision matrix than variables with high variance. We analyze both regimes in order to evaluate the impact of this effect in structure recovery.

In order to measure the closeness of the recovered models to the ground truth, we measured the Kullback-Leibler (KL) divergence, sensitivity (one minus the fraction of falsely excluded edges) and specificity (one minus the fraction of falsely included edges). We compare to the following methods: covariance selection [Banerjee et al., 2006], graphical lasso [Friedman et al., 2007b], Meinshausen-Bühlmann approximation [Meinshausen and Bühlmann, 2006] and Tikhonov regularization. For our method, we found that the variable selection parameter $\tau = 50\rho$ provides reasonable results, in both synthetic and real-world experiments. Therefore, we report results only with respect to the sparseness parameter ρ.

First, we test the performance of our methods for increasing number of variables, moderate edge density (0.5) and high proportion of connected nodes (0.8). Figure 3.1 shows the ROC curves and KL divergence between the recovered models and the ground truth. In both "low" and "high variance confounders" regimes, our $\ell_{1,2}$ and $\ell_{1,\infty}$ methods recover ground truth edges better than competing methods (higher ROC) and produce better probability distributions (lower KL divergence) than the other methods. Our methods degrade less than competing methods in recovering the ground truth edges when the number of variables grows, while the KL divergence behavior remains similar.

Second, we test the performance of our methods with respect to edge density and the proportion of connected nodes. Figure 3.2 shows the KL divergence between the recovered models and the ground truth for the "low variance confounders" regime. Our $\ell_{1,2}$ and $\ell_{1,\infty}$ methods produce better probability distributions (lower KL divergence) than the remaining techniques. (Please, see Appendix D for results on ROC and the "high variance confounders" regime.)

Our $\ell_{1,2}$ method takes 0.07s for $N = 100$, 0.12s for $N = 200$ variables. Our $\ell_{1,\infty}$ method takes 0.13s for $N = 100$, 0.63s for $N = 200$. Graphical lasso [Friedman et al., 2007b], the fastest and most accurate competing method in our evaluation, takes 0.11s for $N = 100$, 0.49s for $N = 200$. Our $\ell_{1,\infty}$ method is slightly slower than graphical lasso, while our $\ell_{1,2}$ method is the fastest. One reason for this is that Lemma 3.5 eliminates more variables in the $\ell_{1,2}$ setting.

For experimental validation on real-world datasets, we use datasets with a diverse nature of probabilistic relationships: brain fMRI, gene expression, NASDAQ stock prices and world weather. The *brain fMRI* dataset collected by Goldstein et al. [2007] captures brain function of 15 cocaine addicted and 11 control subjects under conditions of monetary reward. Each subject contains 87 scans of $53 \times 63 \times 46$ voxels each, taken every 3.5 seconds. Registration to a common spatial template and spatial smoothing was done in SPM2 (`http://www.fil.ion.ucl.ac.uk/spm/`). After sampling each $4 \times 4 \times 4$ voxels, we obtained 869 variables. The *gene expression* dataset contains 8,565 variables and 587 samples. The dataset was collected by Natsoulis et al. [2005] from drug treated rat livers, by treating rats with a variety of fibrate, statin, or estrogen receptor agonist compounds. The dataset is publicly available at `http://www.ebi.ac.uk/`. In order to consider the full set of genes, we had to impute a very small percentage (0.90%) of missing values by randomly generating values with the same mean and standard deviation. The *NASDAQ stocks* dataset contains daily opening and closing prices for 2,749 stocks from Apr 19, 2010 to Apr 18, 2011 (257 days). The dataset was downloaded from `http://www.google.com/finance`. For our experiments, we computed the percentage of change between the closing and opening prices. The *world weather* dataset contains monthly measurements of temperature, precipitation, vapor, cloud cover, wet days and frost days from Jan 1990 to Dec 2002 (156 months) on a 2.5×2.5 degree grid that covers the entire world. The dataset is publicly available at `http://www.cru.uea.ac.uk/`. After sampling each 5×5 degrees, we obtained 4,146 variables. For our experiments, we computed the change between each month and the month in the previous year.

For all the datasets, we used one third of the data for training, one third for validation and the remaining third for testing. Since the brain fMRI dataset has a very small number of subjects, we performed six repetitions by making each third of the data take turns as training, validation and testing sets. In our evaluation, we included scale free networks [Liu and Ihler, 2011]. We did not include the covariance selection method [Banerjee et al., 2006] since we found it is extremely slow for these high-dimensional datasets. We report the negative log-likelihood on the testing set in Figure 3.3 (we subtracted the entropy measured on the testing set and then scaled the results for visualization purposes). We can observe that the log-likelihood of our method is remarkably better than the other techniques for all the datasets.

Regarding comparison to group sparse methods, in our previous experiments we did not include block structure for known block-variable assignments [Duchi et al., 2008a, Schmidt et al., 2009] since our synthetic and real-world datasets lack such assignments. We did not include block structure for unknown assignments [Marlin and K.Murphy, 2009, Marlin et al., 2009] given their time complexity ([Marlin et al., 2009] has a $\mathcal{O}(N^5)$-time Gibbs sampler step for N variables and it is applied for $N = 60$ only, while [Marlin and K.Murphy, 2009] has a $\mathcal{O}(N^4)$-time ridge regression step). Instead, we evaluated our method in the *baker's yeast* gene expression dataset in [Duchi et al., 2008a] which contains 677 variables and 173 samples. We used the experimental settings of Figure 3 in [Marlin and K.Murphy, 2009]. For learning one structure, [Marlin and K.Murphy, 2009] took 5 hours while our $\ell_{1,2}$ method took only 50 seconds. Our method outperforms block structures for known and unknown assignments. The log-likelihood is 0 for Tikhonov regularization, 6 for [Duchi et al., 2008a, Marlin and K.Murphy, 2009], 8 for [Schmidt et al., 2009], and 22 for our $\ell_{1,2}$ method.

We show the structures learnt for cocaine addicted and control subjects in Figure 3.4, for our $\ell_{1,2}$ method and graphical lasso [Friedman et al., 2007b]. The disconnected variables are not shown. Note that our structures involve remarkably fewer connected variables but yield a higher log-likelihood than graphical lasso (Figure 3.3), which suggests that the discarded edges from the disconnected nodes are not important for accurate modeling of this dataset. Moreover, removal of a large number of nuisance variables (voxels) results into a more interpretable model, clearly demonstrating brain areas involved in structural model differences that discriminate cocaine addicted from control subjects. Note that graphical lasso (bottom of Figure 3.4) connects most of the brain voxels in both populations, making them impossible to compare. Our approach produces more "localized" networks (top of the Figure 3.4) involving a relatively small number of brain areas: cocaine addicted subjects show increased interactions between the visual cortex (back of the brain, on the left in the image) and the prefrontal cortex (front of the brain, on the right in the image), while at the same time decreased density of interactions between the visual cortex with other brain areas (more clearly present in control subjects). The alteration in this pathway in the addict group is highly significant from a neuroscientific perspective. First, the trigger for reward was a visual stimulus. Abnormalities in the visual cortex was reported in [Lee et al., 2003] when comparing cocaine abusers to control subjects. Second, the prefrontal cortex is involved in higher-order cognitive functions such as decision making and reward processing. Abnormal monetary processing in the prefrontal cortex was reported in [Goldstein et al., 2009] when comparing cocaine addicted individuals to controls. Although a more careful interpretation of the observed results remains to be done in the near future, these results are encouraging and lend themselves to specific neuroscientific hypothesis testing.

In a different evaluation, we used generatively learnt structures for a classification task. We performed a five-fold cross-validation on the subjects. From the subjects in the training set, we learned one structure for cocaine addicted and one structure for control subjects. Then, we assigned a test subject to the structure that gave highest probability for his data. All methods in our evaluation except Tikhonov regularization obtained 84.6% accuracy. Tikhonov regularization obtained 65.4% accuracy. Therefore, our method produces structures that retain discriminability with respect to standard sparseness promoting techniques.

3.6 Concluding Remarks

In this chapter, we presented variable selection in the context of learning sparse Gaussian graphical models by adding an $\ell_{1,p}$-norm regularization term, for $p \in \{2, \infty\}$. We presented a block coordinate descent method which yields sparse and positive definite estimates. We solved the original problem by efficiently solving a sequence of strictly convex (ℓ_1,ℓ_p) regularized quadratic minimization subproblems.

The motivation behind this work was to incorporate variable selection into structure learning of sparse Markov networks, and specifically Gaussian graphical models. Besides providing a better regularizer (as observed on several real-world datasets: brain fMRI, gene expression, NASDAQ stock prices and world weather), key advantages of our approach include a more accurate structure recovery in the presence of multiple noisy variables (as demonstrated by simulations), significantly better interpretability and same discriminability of the resulting network in practical applications (as shown for brain fMRI analysis).

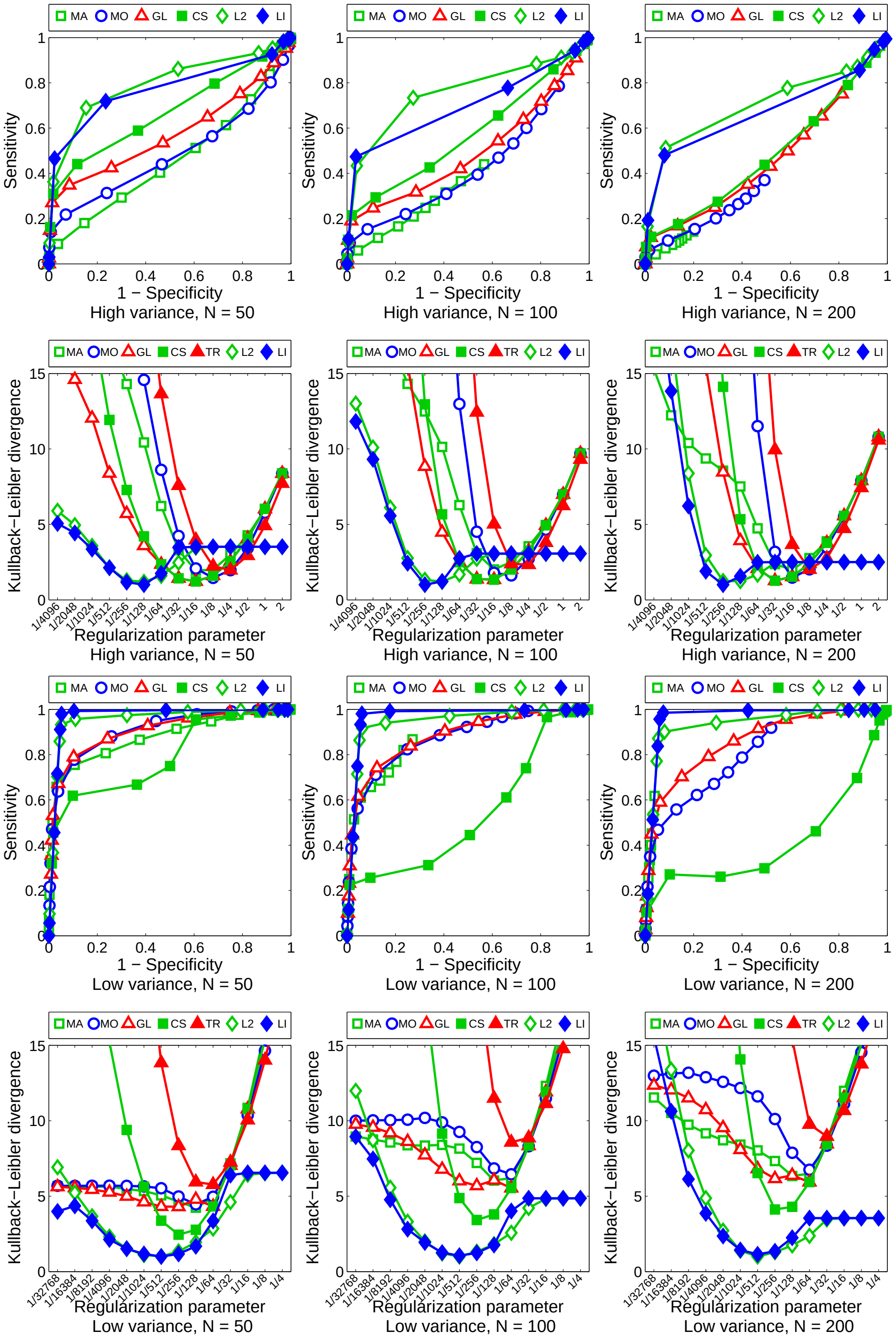


Figure 3.1: ROC curves (first row) and KL divergence (second row) for the "high variance confounders" regime. ROC curves (third row) and KL divergence (fourth row) for the "low variance confounders" regime. Left: $N = 50$ variables, center: $N = 100$ variables, right: $N = 200$ variables (connectedness 0.8, edge density 0.5). Our proposed methods $\ell_{1,2}$ (L2) and $\ell_{1,\infty}$ (LI) recover edges better and produce better probability distributions than Meinshausen-Bühlmann with AND-rule (MA), OR-rule (MO), graphical lasso (GL), covariance selection (CS) and Tikhonov regularization (TR). Our methods degrade less in recovering the ground truth edges when the number of variables grows.

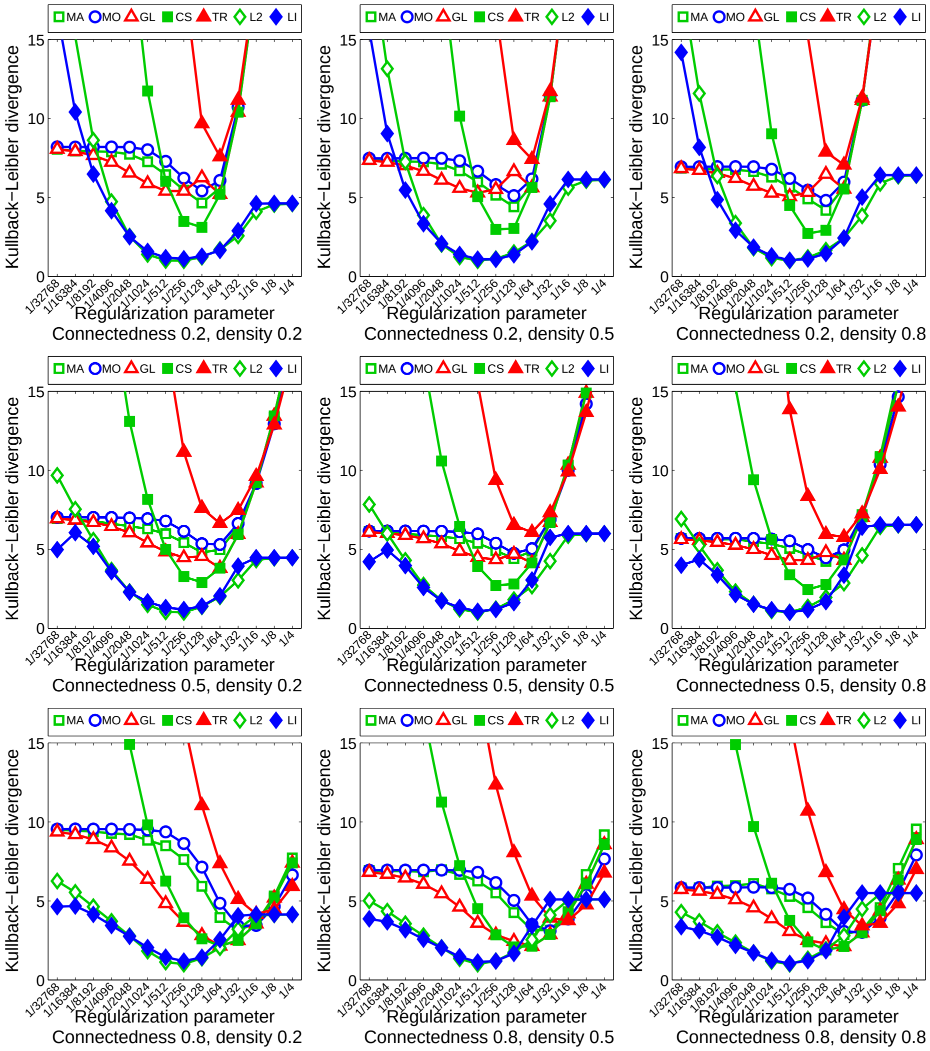
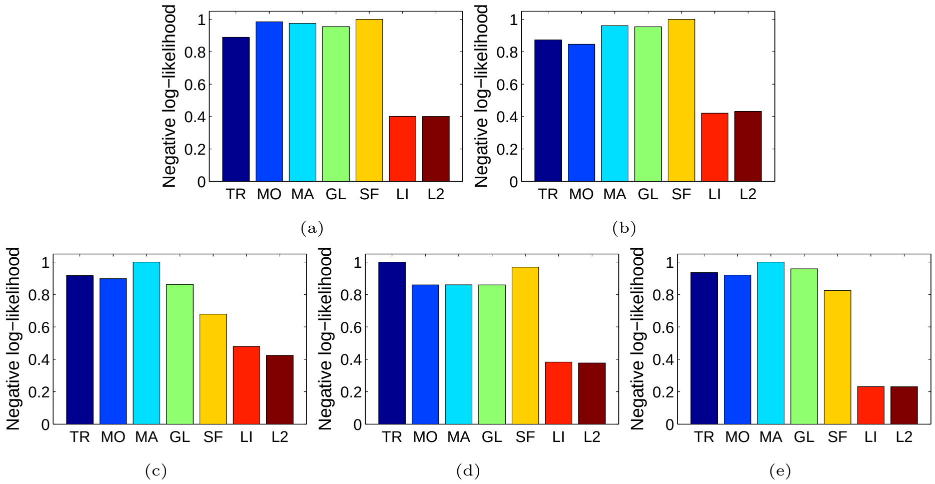


Figure 3.2: Cross-validated KL divergence for structures learnt for the "low variance confounders" regime ($N = 50$ variables, different connectedness and density levels). Our proposed methods $\ell_{1,2}$ (L2) and $\ell_{1,\infty}$ (LI) produce better probability distributions than Meinshausen-Bühlmann with AND-rule (MA), OR-rule (MO), graphical lasso (GL), covariance selection (CS) and Tikhonov regularization (TR).



(a)

(b)

(c)

(d)

(e)

Figure 3.3: Test negative log-likelihood of structures learnt for (a) addicted subjects and (b) control subjects in the brain fMRI dataset, (c) gene expression, (d) NASDAQ stocks and (e) world weather. Our proposed methods $\ell_{1,2}$ (L2) and $\ell_{1,\infty}$ (LI) outperforms the Meinshausen-Bühlmann with AND-rule (MA), OR-rule (MO), graphical lasso (GL), Tikhonov regularization (TR) and scale free networks (SF).

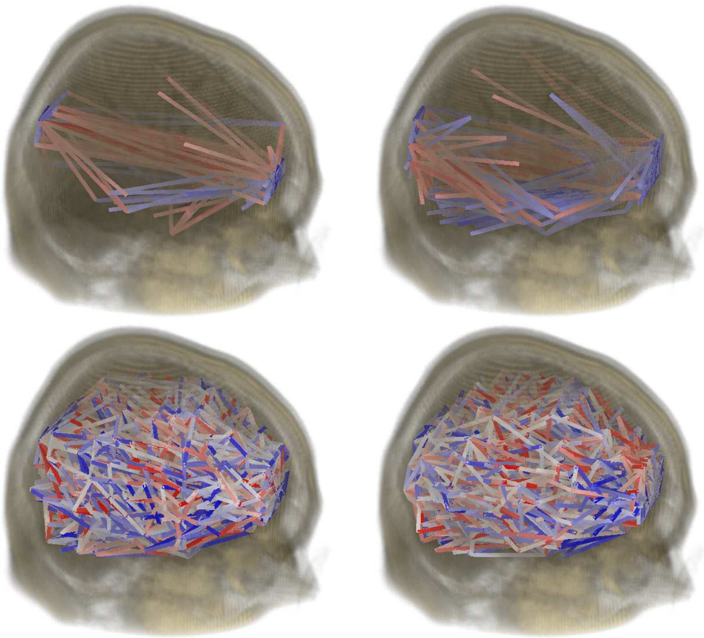


Figure 3.4: Structures learnt for cocaine addicted (left) and control subjects (right), for our $\ell_{1,2}$ method (top) and graphical lasso (bottom). Regularization parameter $\rho = 1/16$. Positive interactions in blue, negative interactions in red. Our structures are sparser (density 0.0016) than graphical lasso (density 0.023) where the number of edges in a complete graph is $\approx$378000.

Chapter 4

Learning Multiple Gaussian MRFs

In the previous chapters, we proposed priors for learning a single structure. In this chapter, we propose priors for the simultaneous learning of multiple structures.

We present $\ell_{1,p}$ *multi-task structure learning* for Gaussian graphical models. We discuss the uniqueness and boundedness of the optimal solution of the maximization problem. A block coordinate descent method leads to a provably convergent algorithm that generates a sequence of positive definite solutions. Thus, we reduce the original problem into a sequence of strictly convex ℓ_p regularized quadratic minimization subproblems. We further show that this subproblem leads to the *continuous quadratic knapsack problem* for $p = \infty$ and to a separable version of the well-known *quadratic trust-region problem* for $p = 2$, for which very efficient methods exist. Finally, we show promising results in synthetic experiments as well as in two real-world datasets.

4.1 Introduction

Structure learning techniques are very useful for analyzing datasets for which probabilistic dependencies are not known apriori. For instance, these techniques allow for modeling interactions between brain regions, based on measured activation levels through imaging. Suppose that we want to learn the structure of brain region interactions for one person. We can expect that the interaction patterns in the brains of two persons are not exactly the same. On the other hand, when learning the structure for one person, we would like to use evidence from other persons as side information in our learning process. This becomes more important in settings with limited amount of data and high variability, such as in functional magnetic resonance image (fMRI) studies. Multi-task learning allows for a more efficient use of training data which is available for multiple related tasks.

In this chapter, we consider the computational aspect of $\ell_{1,p}$ multi-task structure learning, which generalizes the learning of sparse Gaussian graphical models to the multi-task setting by replacing the ℓ_1-norm regularization with an $\ell_{1,p}$-norm, also known as the simultaneous prior [Turlach et al., 2005, Tropp, 2006] for $p = \infty$ or the group-sparse prior [Yuan and Lin, 2006, Meier et al., 2008] for $p = 2$.

Our contribution in this chapter is three-fold. First, we present a block coordinate descent method which is provably convergent and yields sparse and positive definite estimates.

Second, we show the connection between our $\ell_{1,p}$ multi-task structure learning problem and the continuous quadratic knapsack problem for $p = \infty$, which allows us to use existing efficient methods [Helgason et al., 1980, Brucker, 1984, Kiwiel, 2007]. We also show the connection between our multi-task structure learning problem and the quadratic trust-region problem for $p = 2$, which can be efficiently solved by one-dimensional optimization. Third, we discuss penalization of the diagonals of the precision matrices and experimentally show that penalizing the diagonals does not lead to a better generalization performance, when compared to not penalizing the diagonals.

Compared to our short conference version [Honorio and Samaras, 2010], we present a more general framework which assumes $p > 1$, while [Honorio and Samaras, 2010] assumes $p = \infty$. We present a new algorithm for $p = 2$ and experimentally show that our method recovers the ground truth edges and the probability distribution always better than the $\ell_{1,2}$ method of Varoquaux et al. [2010] for every regularization level. We discuss penalization of the diagonals of the precision matrices which leads to additional optimization problems, namely the *continuous logarithmic knapsack problem* for $p = \infty$ and the *separable logarithmic trust-region problem* for $p = 2$. We show that our method outperforms others in recovering the topology of the ground truth model through synthetic experiments. In addition to the small fMRI dataset used in [Honorio and Samaras, 2010], we include validation in a considerably larger fMRI dataset. We experimentally show that the cross-validated log-likelihood of our method is higher than competing methods in both real-world datasets.

Section 4.2 introduces techniques for learning Gaussian graphical models from data. Section 4.3 sets up the $\ell_{1,p}$ multi-task structure learning problem and discusses some of its properties. Section 4.4 describes our block coordinate descent method. Section 4.5 shows the connection to the continuous quadratic knapsack problem. Section 4.6 shows the connection to the quadratic trust-region problem. Section 4.7 presents our algorithm in detail. Section 4.8 discusses penalization of the diagonals of the precision matrices. Experimental results are shown and explained in Section 4.9. Main contributions and results are summarized in Section 4.10.

4.2 Background

In Section 1.4, we introduced Gaussian graphical models as well as techniques for learning sparse Gaussian graphical models through ℓ_1 regularization, such as: *covariance selection* [Banerjee et al., 2006], *graphical lasso* [Friedman et al., 2007b] and the *Meinshausen-Bühlmann approximation* [Meinshausen and Bühlmann, 2006].

Besides sparseness, several regularizers have been proposed for Gaussian graphical models for *single-task* learning, for enforcing diagonal structure [Levina et al., 2008], block structure for known block-variable assignments [Duchi et al., 2008a, Schmidt et al., 2009] and unknown block-variable assignments [Marlin and K.Murphy, 2009, Marlin et al., 2009], spatial coherence [Honorio et al., 2009], sparse changes in controlled experiments [Zhang and Wang, 2010], power law regularization in scale free networks [Liu and Ihler, 2011], or variable selection [Honorio et al., 2012].

Multi-task learning has been applied to very diverse problems, such as linear regression [Liu et al., 2009a,b], classification [Jebara, 2004], compressive sensing [Qi et al., 2008], reinforcement learning [Wilson et al., 2007] and structure learning of Bayesian networks [Niculescu-Mizil and Caruana, 2007].

4.3 Preliminaries

In this section, we set up the problem and discuss some of its properties.

4.3.1 Problem Setup

We propose a prior that is motivated from the multi-task learning literature. Given K arbitrary tasks, our goals are to learn one structure for each task that best explains the observed data, and to promote a common sparseness pattern of edges for all tasks.

For a given task k, we learn a precision matrix $\mathbf{\Omega}^{(k)} \in \mathbb{R}^{N \times N}$ for N variables. Our multi-task regularizer penalizes corresponding edges across tasks (i.e. $\omega_{n_1n_2}^{(1)}, \dots, \omega_{n_1n_2}^{(K)}$). Let $\widehat{\mathbf{\Sigma}}^{(k)} \succeq \mathbf{0}$ be the dense sample covariance matrix for task k, and $T^{(k)} > 0$ be the number of samples in task k. The $\ell_{1,p}$ *multi-task structure learning problem* is defined as:

$$\max_{(\forall k)\ \mathbf{\Omega}^{(k)} \succ \mathbf{0}} \left(\sum_k T^{(k)} (\log\det \mathbf{\Omega}^{(k)} - \langle \widehat{\mathbf{\Sigma}}^{(k)}, \mathbf{\Omega}^{(k)} \rangle) - \rho \|\mathbf{\Omega}\|_{1,p} \right) \tag{4.1}$$

for regularization parameter $\rho > 0$ and $\ell_{1,p}$-norm for $p > 1$. The term $T^{(k)}(\log\det \mathbf{\Omega}^{(k)} - \langle \widehat{\mathbf{\Sigma}}^{(k)}, \mathbf{\Omega}^{(k)} \rangle)$ is the Gaussian log-likelihood for task k, while the term $\|\mathbf{\Omega}\|_{1,p}$ is our multi-task regularizer, and it is defined as:

$$\|\mathbf{\Omega}\|_{1,p} = \sum_{n_1n_2} \|(\omega_{n_1n_2}^{(1)}, \dots, \omega_{n_1n_2}^{(K)})\|_p \tag{4.2}$$

We assume that $p > 1$, since for $p = 1$, the multi-task problem in eq.(4.1) reduces to K *single-task* problems as in eq.(1.1), and for $p < 1$, eq.(4.1) is not convex. The number of samples $T^{(k)}$ is a term that is usually dropped for covariance selection and graphical lasso as in eq.(1.1). For the multi-task structure learning problem, it is important to keep this term when adding the log-likelihood of several tasks into a single objective function.

The $\ell_{1,2}$ multi-task structure learning problem was originally proposed in [Varoquaux et al., 2010], where the authors minimize the original non-smooth objective function by using a sequence of smooth quadratic upper bounds. Varoquaux et al. [2010] do not provide any guarantee of positive definiteness, eigenvalue bounds or convergence. The $\ell_{1,\infty}$ multi-task problem was originally proposed in Honorio and Samaras [2010]. In this chapter, we analyze the computational aspects of the more general $\ell_{1,p}$ multi-task problem for $p > 1$. While the $\ell_{1,\infty}$ multi-task problem of Honorio and Samaras [2010] leads to the *continuous quadratic knapsack problem*, we show that the $\ell_{1,2}$ multi-task problem leads to the *quadratic trust-region problem*. Another multi-task penalty has been proposed in Guo et al. [2010], however this penalty is non-convex.

4.3.2 Bounds

In what follows, we discuss the uniqueness and boundedness of the optimal solution of the multi-task structure learning problem.

Lemma 4.1. *For $\rho > 0$ and $p > 1$, the $\ell_{1,p}$ multi-task structure learning problem in eq.*(4.1) *is a maximization problem with a concave (but not strictly concave) objective function and convex constraints.*

Proof. The Gaussian log-likelihood is concave, since $\log\det$ is concave on the space of symmetric positive definite matrices and $\langle\cdot,\cdot\rangle$ is a linear operator. The multi-task regularizer defined in eq.(4.2) is a non-smooth convex function. Finally, $\mathbf{\Omega}^{(k)} \succ \mathbf{0}$ is a convex constraint. □

Theorem 4.2. *For $\rho > 0$ and $p > 1$, the optimal solution to the $\ell_{1,p}$ multi-task structure learning problem in eq.*(4.1) *is unique and bounded as follows:*

$$(\forall k)\ \left(\frac{1}{\|\widehat{\mathbf{\Sigma}}^{(k)}\|_2 + \frac{N\rho}{T^{(k)}}}\right)\mathbf{I} \preceq \mathbf{\Omega}^{(k)^*} \preceq \left(\frac{NK}{\rho}\right)\mathbf{I} \tag{4.3}$$

Proof. Let the $\ell_{p'}$-norm be the dual of the ℓ_p-norm, i.e. $\frac{1}{p} + \frac{1}{p'} = 1$. By using the identity for dual norms $\rho\|\mathbf{c}\|_p = \max_{\|\mathbf{a}\|_{p'} \leq \rho} \mathbf{a}^{\mathrm{T}}\mathbf{c}$ in eq.(4.1), we get:

$$\max_{(\forall k)\ \mathbf{\Omega}^{(k)} \succ \mathbf{0}}\ \min_{(\forall n_1 n_2)\ \|\mathbf{a}_{n_1 n_2}\|_{p'} \leq \rho} \sum_k T^{(k)} \left(\log\det \mathbf{\Omega}^{(k)} - \langle \widehat{\mathbf{\Sigma}}^{(k)} + \frac{\mathbf{A}^{(k)}}{T^{(k)}}, \mathbf{\Omega}^{(k)} \rangle\right) \tag{4.4}$$

where $\mathbf{a}_{n_1 n_2} = (a^{(1)}_{n_1 n_2}, \ldots, a^{(K)}_{n_1 n_2})^{\mathrm{T}}$ and $\mathbf{A}^{(k)} \in \mathbb{R}^{N \times N}$. By virtue of Sion's minimax theorem, we can swap the order of max and min. Furthermore, note that the optimal solution of the inner equation is independent for each k and is given by $\mathbf{\Omega}^{(k)} = (\widehat{\mathbf{\Sigma}}^{(k)} + \frac{\mathbf{A}^{(k)}}{T^{(k)}})^{-1}$. By replacing this solution in eq.(4.4), we get the dual problem of eq.(4.1):

$$\min_{(\forall n_1 n_2)\ \|\mathbf{a}_{n_1 n_2}\|_{p'} \leq \rho} -\sum_k T^{(k)} \log\det\left(\widehat{\mathbf{\Sigma}}^{(k)} + \frac{\mathbf{A}^{(k)}}{T^{(k)}}\right) - NK \tag{4.5}$$

In order to find a lower bound for the minimum eigenvalue of $\mathbf{\Omega}^{(k)^*}$, note that $\|\mathbf{\Omega}^{(k)^{*-1}}\|_2 = \|\widehat{\mathbf{\Sigma}}^{(k)} + \frac{\mathbf{A}^{(k)}}{T^{(k)}}\|_2 \leq \|\widehat{\mathbf{\Sigma}}^{(k)}\|_2 + \|\frac{\mathbf{A}^{(k)}}{T^{(k)}}\|_2 = \|\widehat{\mathbf{\Sigma}}^{(k)}\|_2 + \frac{1}{T^{(k)}}\|\mathbf{A}^{(k)}\|_2 \leq \|\widehat{\mathbf{\Sigma}}^{(k)}\|_2 + \frac{1}{T^{(k)}}\|\mathbf{A}^{(k)}\|_{\mathfrak{F}}$. Since $\|\mathbf{a}_{n_1 n_2}\|_{p'} \leq \rho$, it follows that $|a^{(k)}_{n_1 n_2}| \leq \rho$ and therefore $\|\mathbf{A}^{(k)}\|_{\mathfrak{F}} \leq N\rho$.

In order to find an upper bound for the maximum eigenvalue of $\mathbf{\Omega}^{(k)^*}$, note that, at optimum, the primal-dual gap is zero:

$$-NK + \sum_k T^{(k)} \langle \widehat{\mathbf{\Sigma}}^{(k)}, \mathbf{\Omega}^{(k)^*} \rangle + \rho\|\mathbf{\Omega}^*\|_{1,p} = 0 \tag{4.6}$$

The upper bound is found as follows: $\|\mathbf{\Omega}^{(k)^*}\|_2 \leq \|\mathbf{\Omega}^{(k)^*}\|_{\mathfrak{F}} \leq \|\mathbf{\Omega}^{(k)^*}\|_1 \leq \|\mathbf{\Omega}^*\|_{1,p} = \frac{NK - \sum_k T^{(k)} \langle \widehat{\mathbf{\Sigma}}^{(k)}, \mathbf{\Omega}^{(k)^*} \rangle}{\rho}$ and since $\widehat{\mathbf{\Sigma}}^{(k)} \succeq \mathbf{0}$ and $\mathbf{\Omega}^{(k)^*} \succ \mathbf{0}$, it follows that $\langle \widehat{\mathbf{\Sigma}}^{(k)}, \mathbf{\Omega}^{(k)^*} \rangle \geq 0$. □

4.4 Block Coordinate Descent Method

In this section, we develop a block coordinate descent method for our $\ell_{1,p}$ multi-task structure learning problem, and discuss some of its properties.

Since the objective function in eq.(4.1) contains a non-smooth regularizer, methods such as gradient descent cannot be applied. On the other hand, subgradient descent methods very rarely converge to non-smooth points [Duchi and Singer, 2009b]. In our problem, these non-smooth points correspond to zeros in the precision matrix, are often the true minima of the objective function, and are very desirable in the solution because they convey information of conditional independence among variables.

We apply block coordinate descent method on the primal problem [Honorio et al., 2009, Honorio and Samaras, 2010, Honorio et al., 2012], unlike covariance selection [Banerjee et al., 2006] and graphical lasso [Friedman et al., 2007b] which optimize the dual. We choose to optimize the primal because the dual formulation in eq.(4.5) leads to a sum of K terms ($\log\det$ functions) which cannot be simplified to a quadratic problem unless $K = 1$.

For clarity of exposition, we will first assume that the diagonals of $\mathbf{\Omega}^{(1)}, \dots, \mathbf{\Omega}^{(K)}$ are not penalized by our multi-task regularizer defined in eq.(4.2). In Section 4.8, we will discuss penalization of the diagonals, for which an additional *continuous logarithmic knapsack problem* for $p = \infty$ or *separable logarithmic trust-region problem* for $p = 2$ needs to be solved. We point out that all the following theorems and lemmas still hold in that case.

Lemma 4.3. *The solution sequence generated by the block coordinate descent method is bounded and every cluster point is a solution of the $\ell_{1,p}$ multi-task structure learning problem in eq.*(4.1).

Proof. The non-smooth regularizer $\|\mathbf{\Omega}\|_{1,p}$ is separable into a sum of $\mathcal{O}(N^2)$ individual functions of the form $\|(\omega^{(1)}_{n_1 n_2}, \dots, \omega^{(K)}_{n_1 n_2})\|_p$. These functions are defined over blocks of K variables, i.e. $\omega^{(1)}_{n_1 n_2}, \dots, \omega^{(K)}_{n_1 n_2}$. The objective function in eq.(4.1) is continuous on a compact level set. By virtue of Theorem 4.1 in Tseng [2001], we prove our claim. □

Theorem 4.4. *The block coordinate descent method for the $\ell_{1,p}$ multi-task structure learning problem in eq.*(4.1) *generates a sequence of positive definite solutions.*

Proof. Maximization can be performed with respect to one row and column of all precision matrices $\mathbf{\Omega}^{(k)}$ at a time. Without loss of generality, we use the last row and column in our derivation, since permutation of rows and columns is always possible. Let:

$$\mathbf{\Omega}^{(k)} = \begin{bmatrix} \mathbf{W}^{(k)} & \mathbf{y}^{(k)} \\ \mathbf{y}^{(k)\mathrm{T}} & z^{(k)} \end{bmatrix}, \widehat{\mathbf{\Sigma}}^{(k)} = \begin{bmatrix} \mathbf{S}^{(k)} & \mathbf{u}^{(k)} \\ \mathbf{u}^{(k)\mathrm{T}} & v^{(k)} \end{bmatrix} \tag{4.7}$$

where $\mathbf{W}^{(k)}, \mathbf{S}^{(k)} \in \mathbb{R}^{N-1\times N-1}$, $\mathbf{y}^{(k)}, \mathbf{u}^{(k)} \in \mathbb{R}^{N-1}$.

In terms of the variables $\mathbf{y}^{(k)}$, $z^{(k)}$ and the constant matrix $\mathbf{W}^{(k)}$, the multi-task structure learning problem in eq.(4.1) can be reformulated as:

$$\max_{(\forall k)\ \mathbf{\Omega}^{(k)}\succ\mathbf{0}} \begin{pmatrix} \sum_k T^{(k)} \left(\log(z^{(k)} - \mathbf{y}^{(k)\mathrm{T}}\mathbf{W}^{(k)^{-1}}\mathbf{y}^{(k)}) - 2\mathbf{u}^{(k)\mathrm{T}}\mathbf{y}^{(k)} - v^{(k)}z^{(k)}\right) \\ -2\rho\sum_n \|(y^{(1)}_n, \dots, y^{(K)}_n)\|_p \end{pmatrix} \tag{4.8}$$

If $\mathbf{\Omega}^{(k)}$ is a symmetric matrix, according to the Haynsworth inertia formula, $\mathbf{\Omega}^{(k)} \succ \mathbf{0}$ if and only if its Schur complement $z^{(k)} - \mathbf{y}^{(k)\mathrm{T}}\mathbf{W}^{(k)^{-1}}\mathbf{y}^{(k)} > 0$ and $\mathbf{W}^{(k)} \succ \mathbf{0}$. By maximizing eq.(4.8) with respect to $z^{(k)}$, we get:

$$z^{(k)} - \mathbf{y}^{(k)\mathrm{T}}\mathbf{W}^{(k)^{-1}}\mathbf{y}^{(k)} = \frac{1}{v^{(k)}} \tag{4.9}$$

and since $v^{(k)} > 0$, this implies that the Schur complement in eq.(4.9) is positive.

Finally, in an iterative optimization algorithm, it suffices to initialize $\mathbf{\Omega}^{(k)}$ to a matrix that is known to be positive definite, e.g. a diagonal matrix with positive elements. □

Remark 4.5. *Note that eq.*(4.9) *defines the "diagonal update step" of the block coordinate descent method. For each k we set $z^{(k)}$ to its optimal value, i.e. $z^{(k)^*} = \frac{1}{v^{(k)}} + \mathbf{y}^{(k)\mathrm{T}}\mathbf{W}^{(k)^{-1}}\mathbf{y}^{(k)}$.*

Theorem 4.6. *The "off-diagonal update step" of the block coordinate descent method for the $\ell_{1,p}$ multi-task structure learning problem in eq.*(4.1) *is equivalent to solving a sequence of strictly convex $\ell_{1,p}$ regularized quadratic subproblems:*

$$\min_{(\forall k)\ \mathbf{y}^{(k)} \in \mathbb{R}^{N-1}} \begin{pmatrix} \sum_k T^{(k)} \left(\frac{1}{2}\mathbf{y}^{(k)\mathrm{T}} v^{(k)} \mathbf{W}^{(k)^{-1}}\mathbf{y}^{(k)} + \mathbf{u}^{(k)\mathrm{T}}\mathbf{y}^{(k)} \right) \\ +\rho \sum_n \|(y_n^{(1)}, \dots, y_n^{(K)})\|_p \end{pmatrix} \tag{4.10}$$

Proof. By replacing the optimal $z^{(k)}$ given by eq.(4.9) into the objective function in eq.(4.8), we get eq.(4.10). Since $\mathbf{W}^{(k)} \succ \mathbf{0} \Rightarrow \mathbf{W}^{(k)^{-1}} \succ \mathbf{0}$, hence eq.(4.10) is strictly convex. □

As we will show in Section 4.8, the Schur complement is still positive when we penalize the diagonals, i.e. $z^{(k)} - \mathbf{y}^{(k)\mathrm{T}}\mathbf{W}^{(k)^{-1}}\mathbf{y}^{(k)} = \xi > 0$. Note that in such case, $\xi \neq \frac{1}{v^{(k)}}$ in contrast to eq.(4.9) but we can still perform the replacement in eq.(4.8), and therefore Theorem 4.6 still holds when penalizing the diagonals.

Lemma 4.7. *Let the $\ell_{p'}$-norm be the dual of the ℓ_p-norm, i.e. $\frac{1}{p} + \frac{1}{p'} = 1$. If the $\ell_{\infty,p'}$ norm $\max_n \|(T^{(1)}u_n^{(1)}, \dots, T^{(K)}u_n^{(K)})\|_{p'} \leq \rho$, the $\ell_{1,p}$ regularized quadratic problem in eq.*(4.10) *has the minimizer $(\forall k)\ \mathbf{y}^{(k)^*} = \mathbf{0}$.*

Proof. The problem in eq.(4.10) has the minimizer $(\forall k)\ \mathbf{y}^{(k)^*} = \mathbf{0}$ if and only if $\mathbf{0}$ belongs to the subdifferential set of the non-smooth objective function at $(\forall k)\ \mathbf{y}^{(k)} = \mathbf{0}$, i.e. $(\exists \mathbf{A} \in \mathbb{R}^{N-1 \times K})\ (T^{(1)}\mathbf{u}^{(1)}, \dots, T^{(K)}\mathbf{u}^{(K)}) + \mathbf{A} = \mathbf{0} \wedge \max_n \|(a_{n1}, \dots, a_{nK})\|_{p'} \leq \rho$. This condition is true for $\max_n \|(T^{(1)}u_n^{(1)}, \dots, T^{(K)}u_n^{(K)})\|_{p'} \leq \rho$. □

Remark 4.8. *By using Lemma 4.7, we can reduce the size of the original problem by removing variables in which this condition holds, since it only depends on the dense sample covariance matrix.*

Theorem 4.9. *The coordinate descent method for the $\ell_{1,p}$ regularized quadratic problem in eq.*(4.10) *is equivalent to solving a sequence of strictly convex ℓ_p regularized separable quadratic subproblems:*

$$\min_{\mathbf{x} \in \mathbb{R}^K} \left(\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{Diag}(\mathbf{q})\mathbf{x} - \mathbf{c}^{\mathrm{T}}\mathbf{x} + \rho\|\mathbf{x}\|_p \right) \tag{4.11}$$

Proof. Without loss of generality, we use the last row and column in our derivation, since permutation of rows and columns is always possible. Let:

$$\mathbf{W}^{(k)^{-1}}=\begin{bmatrix}\mathbf{H}_{11}^{(k)} & \mathbf{h}_{12}^{(k)}\\ \mathbf{h}_{12}^{(k)^{\mathrm{T}}} & h_{22}^{(k)}\end{bmatrix}, \mathbf{y}^{(k)}=\begin{bmatrix}\mathbf{y}_1^{(k)}\\ x_k\end{bmatrix}, \mathbf{u}^{(k)}=\begin{bmatrix}\mathbf{u}_1^{(k)}\\ u_2^{(k)}\end{bmatrix} \tag{4.12}$$

where $\mathbf{H}_{11}^{(k)} \in \mathbb{R}^{N-2\times N-2}$, $\mathbf{h}_{12}^{(k)}, \mathbf{y}_1^{(k)}, \mathbf{u}_1^{(k)} \in \mathbb{R}^{N-2}$.

In terms of the variable $\mathbf{x}$ and the constants $q_k = T^{(k)}v^{(k)}h_{22}^{(k)}$, $c_k = -T^{(k)}(v^{(k)}\mathbf{h}_{12}^{(k)^{\mathrm{T}}}\mathbf{y}_1^{(k)} + u_2^{(k)})$, the $\ell_{1,p}$ regularized quadratic problem in eq.(4.10) can be reformulated as in eq.(4.11). Moreover, since $(\forall k)\ T^{(k)} > 0 \wedge v^{(k)} > 0 \wedge h_{22}^{(k)} > 0 \Rightarrow \mathbf{q} > \mathbf{0}$, and therefore eq.(4.11) is strictly convex. □

4.5 Continuous Quadratic Knapsack Problem

In this section, we show the connection between the multi-task structure learning problem and the continuous quadratic knapsack problem, for which very efficient methods exist.

The continuous quadratic knapsack problem has been solved in several areas. [Helgason et al., 1980] provides an $\mathcal{O}(K \log K)$ algorithm which initially sort the breakpoints. [Brucker, 1984] and later [Kiwiel, 2007] provide deterministic linear-time algorithms by using medians of breakpoint subsets. In the context of machine learning, [Duchi et al., 2008b] provides a randomized linear-time algorithm, while [Liu et al., 2009a] provides an $\mathcal{O}(K \log K)$ algorithm. We point out that [Duchi et al., 2008b, Liu et al., 2009a] assume that the weights of the quadratic term are all equal, i.e. $(\forall k)\ q_k = 1$. In this chapter, we assume arbitrary positive weights, i.e. $(\forall k)\ q_k > 0$.

Theorem 4.10. *For $\mathbf{q} > \mathbf{0}$, $\rho > 0$, $p = \infty$, the ℓ_∞ regularized separable quadratic problem in eq.(4.11) is equivalent to the separable quadratic problem with one ℓ_1 constraint:*

$$\min_{\|\mathbf{r}\|_1 \leq \rho} \left(\frac{1}{2}(\mathbf{r}-\mathbf{c})^{\mathrm{T}}\mathbf{Diag}(\mathbf{q})^{-1}(\mathbf{r}-\mathbf{c})\right) \tag{4.13}$$

Furthermore, their optimal solutions are related by $\mathbf{x}^* = \mathbf{Diag}(\mathbf{q})^{-1}(\mathbf{c}-\mathbf{r}^*)$.

Proof. By Lagrangian duality, the problem in eq.(4.13) is the dual of the problem in eq.(4.11). Furthermore, strong duality holds in this case. □

Remark 4.11. *In eq.(4.13), we can assume that $(\forall k)\ c_k \neq 0$. If $(\exists k)\ c_k = 0$, the partial optimal solution is $r_k^* = 0$, and since this assignment does not affect the constraint, we can safely remove r_k from the optimization problem.*

Remark 4.12. *In what follows, we assume that $\|\mathbf{c}\|_1 > \rho$. If $\|\mathbf{c}\|_1 \leq \rho$, the unconstrained optimal solution of eq.(4.13) is also its optimal solution, since $\mathbf{r}^* = \mathbf{c}$ is inside the feasible region given that $\|\mathbf{r}^*\|_1 \leq \rho$.*

Lemma 4.13. *For $\mathbf{q} > \mathbf{0}$, $(\forall k)\ c_k \neq 0$, $\|\mathbf{c}\|_1 > \rho$, the optimal solution $\mathbf{r}^*$ of the separable quadratic problem with one ℓ_1 constraint in eq.(4.13) belongs to the same orthant as the unconstrained optimal solution $\mathbf{c}$, i.e. $(\forall k)\ r_k^* c_k \geq 0$.*

Proof. We prove this by contradiction. Assume $(\exists k_1)\ r^*_{k_1} c_{k_1} < 0$. Let $\mathbf{r}$ be a vector such that $r_{k_1} = 0$ and $(\forall k_2 \neq k_1)\ r_{k_2} = r^*_{k_2}$. The solution $\mathbf{r}$ is feasible, since $\|\mathbf{r}^*\|_1 \leq \rho$ and $\|\mathbf{r}\|_1 = \|\mathbf{r}^*\|_1 - |r^*_{k_1}| \leq \rho$. The difference in the objective function between $\mathbf{r}^*$ and $\mathbf{r}$ is $\frac{1}{2}(\mathbf{r}^* - \mathbf{c})^{\mathrm{T}}\mathbf{Diag(q)}^{-1}(\mathbf{r}^* - \mathbf{c}) - \frac{1}{2}(\mathbf{r} - \mathbf{c})^{\mathrm{T}}\mathbf{Diag(q)}^{-1}(\mathbf{r} - \mathbf{c}) = \frac{1}{2q_{k_1}}({r^*_{k_1}}^2 - 2c_{k_1}r^*_{k_1}) > \frac{{r^*_{k_1}}^2}{2q_{k_1}} > 0$. Thus, the objective function for $\mathbf{r}$ is smaller than for $\mathbf{r}^*$ (the assumed optimal solution), which is a contradiction. □

Theorem 4.14. *For $\mathbf{q} > \mathbf{0}$, $(\forall k)\ c_k \neq 0$, $\|\mathbf{c}\|_1 > \rho$, the separable quadratic problem with one ℓ_1 constraint in eq.(4.13) is equivalent to the continuous quadratic knapsack problem:*

$$\min_{\substack{\mathbf{g} \geq \mathbf{0} \\ \mathbf{1}^{\mathrm{T}}\mathbf{g} = \rho}} \sum_k \frac{1}{2q_k}(g_k - |c_k|)^2 \tag{4.14}$$

Furthermore, their optimal solutions are related by $(\forall k)\ r^*_k = \operatorname{sgn}(c_k) g^*_k$.

Proof. By invoking Lemma 4.13, we can replace $(\forall k)\ r_k = \operatorname{sgn}(c_k) g_k,\ g_k \geq 0$ in eq.(4.13). Finally, we change the inequality constraint $\mathbf{1}^{\mathrm{T}}\mathbf{g} \leq \rho$ to an equality constraint since $\|\mathbf{c}\|_1 > \rho$ and therefore, the optimal solution must be on the boundary of the constraint set. □

Lemma 4.15. *For $\mathbf{q} > \mathbf{0}$, $(\forall k)\ c_k \neq 0$, $\|\mathbf{c}\|_1 > \rho$, the continuous quadratic knapsack problem in eq.(4.14) has the solution:*

$$g_k(\nu) = \max(0, |c_k| - \nu q_k) \tag{4.15}$$

for some ν, and furthermore, the optimal solution fulfills the condition:

$$\mathbf{g}^* = \mathbf{g}(\nu) \Leftrightarrow \mathbf{1}^{\mathrm{T}}\mathbf{g}(\nu) = \rho \tag{4.16}$$

Proof. The Lagrangian of eq.(4.14) is:

$$\min_{\mathbf{g} \geq \mathbf{0}} \left(\sum_k \frac{1}{2q_k}(g_k - |c_k|)^2 + \nu(\mathbf{1}^{\mathrm{T}}\mathbf{g} - \rho) \right) \tag{4.17}$$

Both results can be obtained by invoking the Karush-Kuhn-Tucker optimality conditions on eq.(4.17). □

Remark 4.16. *Note that $g_k(\nu)$ in eq.(4.15) is a decreasing piecewise linear function with breakpoint $\nu = \frac{|c_k|}{q_k} > 0$. By Lemma 4.15, finding the optimal $\mathbf{g}^*$ is equivalent to finding ν in a piecewise linear function $\mathbf{1}^{\mathrm{T}}\mathbf{g}(\nu)$ that produces ρ.*

Lemma 4.17. *For $\mathbf{q} > \mathbf{0}$, $(\forall k)\ c_k \neq 0$, $\|\mathbf{c}\|_1 > \rho$, the continuous quadratic knapsack problem in eq.(4.14) has the optimal solution $g^*_k = \max(0, |c_k| - \nu^* q_k)$ for:*

$$\frac{|c_{\pi_{k^*}}|}{q_{\pi_{k^*}}} \geq \nu^* = \frac{\sum_{k=1}^{k^*} |c_{\pi_k}| - \rho}{\sum_{k=1}^{k^*} q_{\pi_k}} \geq \frac{|c_{\pi_{k^*+1}}|}{q_{\pi_{k^*+1}}} \tag{4.18}$$

where the breakpoints are sorted in decreasing order by a permutation π of the indices $1, 2, \ldots, K$, i.e. $\frac{|c_{\pi_1}|}{q_{\pi_1}} \geq \frac{|c_{\pi_2}|}{q_{\pi_2}} \geq \cdots \geq \frac{|c_{\pi_K}|}{q_{\pi_K}} \geq \frac{|c_{\pi_{K+1}}|}{q_{\pi_{K+1}}} \equiv 0$.

Proof. Given k^*, ν^* can be found straightforwardly by using the equation of the line. In order to find k^*, we search for the range in which $\mathbf{1}^{\mathrm{T}}\mathbf{g}\left(\frac{|c_{\pi_{k^*}}|}{q_{\pi_{k^*}}}\right) \leq \rho \leq \mathbf{1}^{\mathrm{T}}\mathbf{g}\left(\frac{|c_{\pi_{k^*+1}}|}{q_{\pi_{k^*+1}}}\right)$. □

Theorem 4.18. *For $\mathbf{q} > \mathbf{0}$, $\rho > 0$, $p = \infty$, the ℓ_∞ regularized separable quadratic problem in eq.*(4.11) *has the optimal solution:*

$$\begin{array}{l} \|\mathbf{c}\|_1 \leq \rho \Rightarrow \mathbf{x}^* = \mathbf{0} \\ \|\mathbf{c}\|_1 > \rho \wedge k > k^* \Rightarrow x^*_{\pi_k} = \frac{c_{\pi_k}}{q_{\pi_k}} \\ \|\mathbf{c}\|_1 > \rho \wedge k \leq k^* \Rightarrow x^*_{\pi_k} = \mathrm{sgn}(c_{\pi_k})\frac{\sum_{k=1}^{k^*}|c_{\pi_k}|-\rho}{\sum_{k=1}^{k^*} q_{\pi_k}} \end{array} \tag{4.19}$$

Proof. For $\|\mathbf{c}\|_1 \leq \rho$, from Remark 4.12 we know that $\mathbf{r}^* = \mathbf{c}$. By Theorem 4.10, the optimal solution of eq.(4.11) is $\mathbf{x}^* = \mathbf{Diag}(\mathbf{q})^{-1}(\mathbf{c} - \mathbf{r}^*) = \mathbf{0}$, and we prove the first claim.

For $\|\mathbf{c}\|_1 > \rho$, by Theorem 4.10, the optimal solution of eq.(4.11) $x^*_{\pi_k} = \frac{1}{q_{\pi_k}}(c_{\pi_k} - r^*_{\pi_k})$. By Theorem 4.14, $x^*_{\pi_k} = \frac{1}{q_{\pi_k}}(c_{\pi_k} - \mathrm{sgn}(c_{\pi_k})g^*_{\pi_k})$. By Lemma 4.17, $x^*_{\pi_k} = \frac{c_{\pi_k}}{q_{\pi_k}} - \mathrm{sgn}(c_{\pi_k})\max(0, \frac{|c_{\pi_k}|}{q_{\pi_k}} - \nu^*)$.

If $k > k^* \Rightarrow \frac{|c_{\pi_k}|}{q_{\pi_k}} < \nu^* \Rightarrow x^*_{\pi_k} = \frac{c_{\pi_k}}{q_{\pi_k}}$, and we prove the second claim.

If $k \leq k^* \Rightarrow \frac{|c_{\pi_k}|}{q_{\pi_k}} \geq \nu^* \Rightarrow x^*_{\pi_k} = \mathrm{sgn}(c_{\pi_k})\nu^*$, and we prove the third claim. □

4.6 Separable Quadratic Trust-Region Problem

In this section, we show the connection between the $\ell_{1,2}$ multi-task structure learning problem and the separable quadratic trust-region problem, which can be efficiently solved by one-dimensional optimization.

The trust-region problem has been extensively studied by the mathematical optimization community [Forsythe and Golub, 1965, Moré and Sorensen, 1983, Boyd and Vandenberghe, 2006]. Trust-region methods arise in the optimization of general convex functions. In that context, the strategy behind trust-region methods is to perform a local second-order approximation to the original objective function. The quadratic model for local optimization is "trusted" to be correct inside a circular region (i.e. the trust region). Separability is usually not assumed, i.e. a symmetric matrix $\mathbf{Q}$ is used instead of $\mathbf{Diag}(\mathbf{q})$ in eq.(4.11), and therefore the general algorithms are more involved than ours. In the context of machine learning, [Duchi and Singer, 2009b] provides a closed form solution for the separable version of the problem when the weights of the quadratic term are all equal, i.e. $(\forall k)\ q_k = 1$. In this chapter, we assume arbitrary positive weights, i.e. $(\forall k)\ q_k > 0$. A closed form solution is not possible in this general case, but the efficient one-dimensional *Newton-Raphson method* can be applied.

Theorem 4.19. *For $\mathbf{q} > \mathbf{0}$, $\rho > 0$, $p = 2$, the ℓ_2 regularized separable quadratic problem in eq.*(4.11) *is equivalent to the separable quadratic trust-region problem:*

$$\min_{\|\mathbf{r}\|_2 \leq \rho} \left(\frac{1}{2}(\mathbf{r} - \mathbf{c})^{\mathrm{T}}\mathbf{Diag}(\mathbf{q})^{-1}(\mathbf{r} - \mathbf{c})\right) \tag{4.20}$$

Furthermore, their optimal solutions are related by $\mathbf{x}^* = \mathbf{Diag}(\mathbf{q})^{-1}(\mathbf{c} - \mathbf{r}^*)$.

Proof. By Lagrangian duality, the problem in eq.(4.20) is the dual of the problem in eq.(4.11). Furthermore, strong duality holds in this case. □

Remark 4.20. *In eq.*(4.20)*, we can assume that* $(\forall k)\ c_k \neq 0$*. If* $(\exists k)\ c_k = 0$*, the partial optimal solution is* $r_k^* = 0$*, and since this assignment does not affect the constraint, we can safely remove* r_k *from the optimization problem.*

Remark 4.21. *In what follows, we assume that* $\|\mathbf{c}\|_2 > \rho$*. If* $\|\mathbf{c}\|_2 \leq \rho$*, the unconstrained optimal solution of eq.*(4.20) *is also its optimal solution, since* $\mathbf{r}^* = \mathbf{c}$ *is inside the feasible region given that* $\|\mathbf{r}^*\|_2 \leq \rho$*.*

Lemma 4.22. *For* $\mathbf{q} > \mathbf{0}$*,* $(\forall k)\ c_k \neq 0$*,* $\|\mathbf{c}\|_2 > \rho$*, the separable quadratic trust-region problem in eq.*(4.20) *is equivalent to the problem:*

$$\min_{\lambda \geq 0} \left(\sum_n \frac{c_n^2}{q_n + \lambda q_n^2} + \rho^2 \lambda \right) \tag{4.21}$$

Furthermore, their optimal solutions are related by $\mathbf{r}^* = \mathbf{Diag}(\mathbf{1} + \lambda^* \mathbf{q})^{-1} \mathbf{c}$*.*

Proof. By Lagrangian duality, the problem in eq.(4.21) is the dual of the problem in eq.(4.20). Furthermore, strong duality holds in this case. □

Corollary 4.23. *For the special case* $\mathbf{q} = \mathbf{1}$ *of Duchi and Singer [2009b], the trust-region dual problem in eq.*(4.21) *has the closed form solution* $\lambda^* = \max\left(0, \frac{\|\mathbf{c}\|_2}{\rho} - 1\right)$*.*

Proof. For $\mathbf{q} = \mathbf{1}$, the problem in eq.(4.21) becomes $\min_{\lambda \geq 0} \left(\frac{\|\mathbf{c}\|_2^2}{1+\lambda} + \rho^2 \lambda \right)$. By minimizing with respect to λ and by noting that $\lambda \geq 0$, we prove our claim. □

Theorem 4.24. *For* $\mathbf{q} > \mathbf{0}$*,* $\rho > 0$*,* $p = 2$*, the* ℓ_2 *regularized separable quadratic problem in eq.*(4.11) *has the optimal solution:*

$$\begin{aligned} &\|\mathbf{c}\|_2 \leq \rho \Rightarrow \mathbf{x}^* = \mathbf{0} \\ &\|\mathbf{c}\|_2 > \rho \Rightarrow \mathbf{x}^* = \lambda^* \mathbf{Diag}(\mathbf{1} + \lambda^* \mathbf{q})^{-1} \mathbf{c} \end{aligned} \tag{4.22}$$

Proof. For $\|\mathbf{c}\|_2 \leq \rho$, from Remark 4.21 we know that $\mathbf{r}^* = \mathbf{c}$. By Theorem 4.19, the optimal solution of eq.(4.11) is $\mathbf{x}^* = \mathbf{Diag}(\mathbf{q})^{-1}(\mathbf{c} - \mathbf{r}^*) = \mathbf{0}$, and we prove the first claim.

For $\|\mathbf{c}\|_2 > \rho$, by Theorem 4.19, the optimal solution of eq.(4.11) is $(\forall k)\ x_k^* = \frac{1}{q_k}(c_k - r_k^*)$. By Lemma 4.22, $x_k^* = \frac{1}{q_k}(c_k - \frac{1}{1+\lambda^* q_k} c_k) = \frac{\lambda^*}{1+\lambda^* q_k} c_k$, and we prove the second claim. □

4.7 Algorithm

Algorithm 4.1 shows the block coordinate descent method in detail. A careful implementation of the algorithm allows obtaining a time complexity of $\mathcal{O}(LN^3K)$ for L iterations, N variables and K tasks. In our experiments, the algorithm converges quickly in usually $L = 10$ iterations. The polynomial dependence $\mathcal{O}(N^3)$ on the number of variables is expected since we cannot produce an algorithm faster than computing the inverse of the sample covariance in the case of an infinite sample. For $p = \infty$, the linear-time dependence $\mathcal{O}(K)$ on the number of tasks can be accomplished by using a deterministic linear-time method for solving the continuous quadratic knapsack problem, based on medians of breakpoint subsets [Kiwiel, 2007]. A very easy-to-implement $\mathcal{O}(K \log K)$ algorithm is obtained by initially sorting the breakpoints and searching the range for which Lemma 4.17 holds. For $p = 2$, the linear-time dependence $\mathcal{O}(K)$ on the number of tasks can be accomplished by using the one-dimensional Newton-Raphson method for solving the trust-region dual problem in eq.(4.21). In our implementation, we initialize $\lambda = 0$ and perform 10 iterations of the Newton-Raphson method.

Algorithm 4.1 Block Coordinate Descent for Multi-task Learning.

```
Input: $\rho > 0$, for each $k$, $\widehat{\mathbf{\Sigma}}^{(k)} \succeq \mathbf{0}$, $T^{(k)} > 0$
Initialize for each $k$, $\mathbf{\Omega}^{(k)} = \mathbf{Diag}(\widehat{\mathbf{\Sigma}}^{(k)})^{-1}$
for each iteration $1, \dots, L$ and each variable $1, \dots, N$ do
  Split for each $k$, $\mathbf{\Omega}^{(k)}$ into $\mathbf{W}^{(k)}, \mathbf{y}^{(k)}, z^{(k)}$ and $\widehat{\mathbf{\Sigma}}^{(k)}$ into $\mathbf{S}^{(k)}, \mathbf{u}^{(k)}, v^{(k)}$ as described in eq.(4.7)
  Update for each $k$, ${\mathbf{W}^{(k)}}^{-1}$ by using the Sherman-Woodbury-Morrison formula (Note that when iterating
  from one variable to the next one, only one row and column change on matrix $\mathbf{W}^{(k)}$, see Appendix B)
  for each variable $1, \dots, N-1$ do
    Split for each $k$, ${\mathbf{W}^{(k)}}^{-1}, \mathbf{y}^{(k)}, \mathbf{u}^{(k)}$ as in eq.(4.12)
    For $p = \infty$, solve the $\ell_\infty$ regularized separable quadratic problem by eq.(4.19), either by sorting the
    breakpoints or using medians of breakpoint subsets. For $p = 2$, solve the $\ell_2$ regularized separable
    quadratic problem by eq.(4.22) by using the Newton-Raphson method for solving the trust-region
    dual problem in eq.(4.21)
  end for
  Update for each $k$, $z^{(k)} \leftarrow \frac{1}{v^{(k)}} + {\mathbf{y}^{(k)}}^{\mathrm{T}} {\mathbf{W}^{(k)}}^{-1} \mathbf{y}^{(k)}$
end for
Output: for each $k$, $\mathbf{\Omega}^{(k)} \succ \mathbf{0}$
```

4.8 Penalizing the Diagonals

In this section, we discuss penalization of the diagonals of the precision matrices. It is unclear whether diagonal penalization leads to better models with respect to structure as well as generalization performance. For the *single-task* problem, covariance selection [Banerjee et al., 2006] and graphical lasso [Friedman et al., 2007b] penalize the weights of the diagonal elements. In contrast, the analysis of consistency in structure recovery of Ravikumar et al. [2008] assumed that diagonals are not penalized.

Note that, when the diagonals are not penalized, the "diagonal update step" (Remark 4.5) reduces to setting for each k, ${z^{(k)}}^* = \frac{1}{v^{(k)}} + {\mathbf{y}^{(k)}}^{\mathrm{T}} {\mathbf{W}^{(k)}}^{-1} \mathbf{y}^{(k)}$. Penalization of the diagonals of the precision matrices is more involved, since it requires the solution of additional optimization problems, namely the *continuous logarithmic knapsack problem* for $p = \infty$ and the *separable logarithmic trust-region problem* for $p = 2$. First, we discuss the general problem for arbitrary $p > 1$.

Lemma 4.25. *When penalizing the diagonals of the precision matrices, the "diagonal update step" of the block coordinate descent method for the $\ell_{1,p}$ multi-task structure learning prob-*

lem in eq.(4.1) *is equivalent to solving a sequence of strictly convex ℓ_p regularized separable logarithmic subproblems:*

$$\max_{(\forall k)\ z^{(k)}>b_k} \left(\sum_k q_k \log(z^{(k)} - b_k) - \mathbf{c}^{\mathrm{T}}\mathbf{z} - \rho\|\mathbf{z}\|_p \right) \tag{4.23}$$

where $\mathbf{z} = \left(z^{(1)}, \dots, z^{(K)}\right)^{\mathrm{T}}$ and $\mathbf{q}, \mathbf{c}, \mathbf{b} > \mathbf{0}$. Moreover, the block coordinate descent method generates a sequence of positive definite solutions.

Proof. When we choose to penalize the diagonals of the precision matrices $\mathbf{\Omega}^{(1)}, \dots, \mathbf{\Omega}^{(K)}$, eq.(4.8) contains an additional ℓ_p penalty, i.e. $\rho\|\mathbf{z}\|_p$. In terms of the variables $\mathbf{y}^{(k)}, z^{(k)}$ and the constant matrix $\mathbf{W}^{(k)}$ introduced in eq.(4.7), the multi-task structure learning problem in eq.(4.1) can be reformulated as:

$$\max_{(\forall k)\ \mathbf{\Omega}^{(k)}\succ\mathbf{0}} \left(\begin{array}{l} \sum_k T^{(k)} \left(\log(z^{(k)} - \mathbf{y}^{(k)\mathrm{T}} \mathbf{W}^{(k)^{-1}} \mathbf{y}^{(k)}) - 2\mathbf{u}^{(k)\mathrm{T}} \mathbf{y}^{(k)} - v^{(k)} z^{(k)} \right) \\ -2\rho \sum_n \|(y_n^{(1)}, \dots, y_n^{(K)})\|_p - \rho\|\mathbf{z}\|_p \end{array} \right) \tag{4.24}$$

Let $q_k = T^{(k)} > 0$, $c_k = T^{(k)} v^{(k)} > 0$ and $b_k = \mathbf{y}^{(k)\mathrm{T}} \mathbf{W}^{(k)^{-1}} \mathbf{y}^{(k)} \geq 0$ since $\mathbf{W}^{(k)} \succ \mathbf{0}$ and $\mathbf{y}^{(k)}$ is an arbitrary vector (including the case $\mathbf{y}^{(k)} = 0$). We obtain eq.(4.23) by noting that we are maximizing with respect to $\mathbf{z}$ and by enforcing $(\forall k)\ z^{(k)} > b_k$ since $\log(z^{(k)} - b_k)$ is undefined for $z^{(k)} \leq b_k$.

If $\mathbf{\Omega}^{(k)}$ is a symmetric matrix, according to the Haynsworth inertia formula, $\mathbf{\Omega}^{(k)} \succ \mathbf{0}$ if and only if its Schur complement $z^{(k)} - \mathbf{y}^{(k)\mathrm{T}} \mathbf{W}^{(k)^{-1}} \mathbf{y}^{(k)} > 0$ and $\mathbf{W}^{(k)} \succ \mathbf{0}$. Note that the Schur complement $z^{(k)} - \mathbf{y}^{(k)\mathrm{T}} \mathbf{W}^{(k)^{-1}} \mathbf{y}^{(k)} = z^{(k)} - b_k$ and therefore it is strictly positive for every feasible solution given the constraints $(\forall k)\ z^{(k)} > b_k$.

Finally, in an iterative optimization algorithm, it suffices to initialize $\mathbf{\Omega}^{(k)}$ to a matrix that is known to be positive definite, e.g. a diagonal matrix with positive elements. □

Lemma 4.26. *For $\mathbf{q}, \mathbf{c}, \mathbf{b} > \mathbf{0}$, $\rho > 0$, $p > 1$, the ℓ_p regularized separable logarithmic problem in eq.*(4.23) *is equivalent to the separable logarithmic problem with one $\ell_{p'}$ constraint:*

$$\min_{\substack{\mathbf{r}\geq\mathbf{0} \\ \|\mathbf{r}\|_{p'}=\rho}} \left(-\sum_k q_k \log(r_k + c_k) - \mathbf{b}^{\mathrm{T}}\mathbf{r} \right) \tag{4.25}$$

Furthermore, their optimal solutions are related by $z^{(k)^*} = b_k + \frac{q_k}{c_k + r_k^*}$.

Proof. By Lagrangian duality, the problem in eq.(4.25) is the dual of the problem in eq.(4.23). Furthermore, strong duality holds in this case.

The constraint $\mathbf{r} \geq \mathbf{0}$ comes from the fact that $\mathbf{z} > \mathbf{0}$ since it is the diagonal of positive definite matrices. Note that for a general $\mathbf{z} \in \mathbb{R}^K$, we have $\rho\|\mathbf{z}\|_p = \max_{\|\mathbf{r}\|_{p'}\leq\rho} \mathbf{r}^{\mathrm{T}}\mathbf{z}$. In order to maximize this expression, $\mathbf{r}$ will take values on the non-negative orthant since $\mathbf{z} > \mathbf{0}$.

We changed the inequality constraint $\|\mathbf{r}\|_{p'} \leq \rho$ to an equality constraint since the objective is separable and decreasing with respect to each r_k and therefore, the optimal solution must be on the boundary of the constraint set. □

In what follows, we focus on the case $p = \infty$ and show that this problem can be solved by a combination of sorting and the Newton-Raphson method.

Lemma 4.27. *For $\mathbf{q}, \mathbf{c}, \mathbf{b} > \mathbf{0}$, $\rho > 0$, $p = \infty$, the separable logarithmic problem with one $\ell_{p'}$ in eq.(4.25) is the continuous logarithmic knapsack problem:*

$$\min_{\substack{\mathbf{r}\geq\mathbf{0}\\ \mathbf{1}^{\mathrm{T}}\mathbf{r}=\rho}} \left(-\sum_k q_k \log(r_k + c_k) - \mathbf{b}^{\mathrm{T}}\mathbf{r} \right) \tag{4.26}$$

which has the solution:

$$r_k(\nu) = \begin{cases} +\infty, & \nu \leq b_k \\ \frac{q_k}{\nu - b_k} - c_k, & b_k < \nu < \frac{q_k}{c_k} + b_k \\ 0, & \nu \geq \frac{q_k}{c_k} + b_k \end{cases} \tag{4.27}$$

for some ν, and furthermore, the optimal solution fulfills the condition:

$$\mathbf{r}^* = \mathbf{r}(\nu) \Leftrightarrow \mathbf{1}^{\mathrm{T}}\mathbf{r}(\nu) = \rho \tag{4.28}$$

Proof. The Lagrangian of eq.(4.26) is:

$$\min_{\mathbf{r}\geq\mathbf{0}} \left(-\sum_k q_k \log(r_k + c_k) - \mathbf{b}^{\mathrm{T}}\mathbf{r} + \nu(\mathbf{1}^{\mathrm{T}}\mathbf{r} - \rho) \right) \tag{4.29}$$

Both results can be obtained by invoking the Karush-Kuhn-Tucker optimality conditions on eq.(4.29). □

Remark 4.28. *Note that for $\nu \leq b_k$, we have $r_k(\nu) = +\infty$ in eq.(4.27), therefore $\mathbf{1}^{\mathrm{T}}\mathbf{r}(\nu)$ is finite if and only if $\nu > \max_k b_k = \|b\|_\infty$. Additionally, for $\nu > \|b\|_\infty$ we have that $r_k(\nu)$ in eq.(4.27) is a decreasing piecewise inverse function with breakpoint $\nu = \frac{q_k}{c_k} + b_k > 0$. By Lemma 4.27, finding the optimal $\mathbf{r}^*$ is equivalent to finding ν in a piecewise inverse function $\mathbf{1}^{\mathrm{T}}\mathbf{r}(\nu)$ that produces ρ.*

Similarly as in Lemma 4.17, we sort the breakpoints in decreasing order, i.e. we find a permutation π of the indices $1, 2, \ldots, K$ such that $\frac{q_{\pi_1}}{c_{\pi_1}} + b_{\pi_1} \geq \frac{q_{\pi_2}}{c_{\pi_2}} + b_{\pi_2} \geq \cdots \geq \frac{q_{\pi_K}}{c_{\pi_K}} + b_{\pi_K} \geq \frac{q_{\pi_{K+1}}}{c_{\pi_{K+1}}} + b_{\pi_{K+1}} \equiv 0$. Then, we search for the optimal breakpoint k^* or equivalently we search for the range in which $\mathbf{1}^{\mathrm{T}}\mathbf{r}\left(\frac{q_{\pi_{k^*}}}{c_{\pi_{k^*}}} + b_{\pi_{k^*}}\right) \leq \rho \leq \mathbf{1}^{\mathrm{T}}\mathbf{r}\left(\frac{q_{\pi_{k^*+1}}}{c_{\pi_{k^*+1}}} + b_{\pi_{k^*+1}}\right)$. After finding k^*, ν^* can be found by the Newton-Raphson method in order to fulfill the condition $\mathbf{1}^{\mathrm{T}}\mathbf{r}(\nu^*) = \rho$ in Lemma 4.27. In our implementation, we initialize ν at one of the extremes of the optimal range, i.e. $\nu = \max(\|b\|_\infty + \varepsilon, \frac{q_{\pi_{k^*}}}{c_{\pi_{k^*}}} + b_{\pi_{k^*}})$ for some small $\varepsilon > 0$. We then perform 10 iterations of the Newton-Raphson method.

Next, we focus on the case $p = 2$ and show that this problem can be solved by the Newton-Raphson method.

Lemma 4.29. *For $\mathbf{q}, \mathbf{c}, \mathbf{b} > \mathbf{0}$, $\rho > 0$, $p = 2$, the separable logarithmic problem with one $\ell_{p'}$ in eq.*(4.25) *is the separable logarithmic trust-region problem:*

$$\min_{\substack{\mathbf{r} \geq \mathbf{0} \\ \|\mathbf{r}\|_2 \leq \rho}} \left(-\sum_k q_k \log(r_k + c_k) - \mathbf{b}^{\mathrm{T}} \mathbf{r} \right) \tag{4.30}$$

which can be solved by one-dimensional optimization of:

$$\max_{\lambda \geq 0} \left(-\sum_k q_k \log(r_k(\lambda) + c_k) - \mathbf{b}^{\mathrm{T}} \mathbf{r}(\lambda) + \frac{\lambda}{2} \left(\mathbf{r}(\lambda)^{\mathrm{T}} \mathbf{r}(\lambda) - \rho^2 \right) \right) \tag{4.31}$$

where $r_k(\lambda) = \frac{b_k - \lambda c_k + \sqrt{(b_k + \lambda c_k)^2 + 4\lambda q_k}}{2\lambda}$.

Proof. By Lagrangian duality, the problem in eq.(4.31) is the dual of the problem in eq.(4.30). Furthermore, strong duality holds in this case. □

In our implementation, we initialize $\lambda = \frac{1}{K} \sum_k \frac{q_k + b_k(c_k + \rho)}{\rho(c_k + \rho)}$ and perform 10 iterations of the Newton-Raphson method. Our initialization rule follows from using the average of the k independently optimal values of λ. That is, we consider only one task k at a time from eq.(4.31) which leads to $\max_{\lambda \geq 0} \left(-q_k \log(r_k(\lambda) + c_k) - b_k r_k(\lambda) + \frac{\lambda}{2}(r_k^2(\lambda) - \rho^2) \right)$. Then, we compute the optimal value of λ under this setting, which is $\frac{q_k + b_k(c_k + \rho)}{\rho(c_k + \rho)}$. Finally, we average these optimal values for all k which leads to our initialization rule for λ.

4.9 Experimental Results

We begin with a synthetic example to test the ability of the method to recover the ground truth structure from data. The model contains $N = 50$ variables and $K = 5$ tasks. For each of 50 repetitions, we generate a topology (undirected graph) $\mathbf{\Upsilon}_g \in \{0, 1\}^{N \times N}$ with a required edge density (either 0.1,0.3,0.5). For each task k, we first generate a Gaussian graphical model $\mathbf{\Omega}_g^{(k)}$ with topology $\mathbf{\Upsilon}_g$ where each edge weight is generated uniformly at random from $[-1; +1]$. We ensure positive definiteness of $\mathbf{\Omega}_g^{(k)}$ by verifying that its minimum eigenvalue is at least 0.1. We then generate a dataset of $T^{(k)} = 50$ samples.

In order to measure the closeness of the recovered models to the ground truth, we measured the Kullback-Leibler divergence, sensitivity (one minus the fraction of falsely excluded edges) and specificity (one minus the fraction of falsely included edges). For comparison purposes, we used the following *single-task* methods: covariance selection [Banerjee et al., 2006], graphical lasso [Friedman et al., 2007b], Meinshausen-Bühlmann approximation [Meinshausen and Bühlmann, 2006] and Tikhonov regularization. We also compared our method to the $\ell_{1,2}$ multi-task upper bound method of Varoquaux et al. [2010].

Figure 4.1 shows the ROC curves and Kullback-Leibler divergence between the recovered models and the ground truth. Note that both our $\ell_{1,\infty}$ and $\ell_{1,2}$ multi-task methods recover the ground truth edges remarkably better (higher ROC) than the comparison methods, including the $\ell_{1,2}$ multi-task upper bound method of Varoquaux et al. [2010]. Our $\ell_{1,2}$ method always produces better probability distributions (lower Kullback-Leibler divergence) than the $\ell_{1,2}$

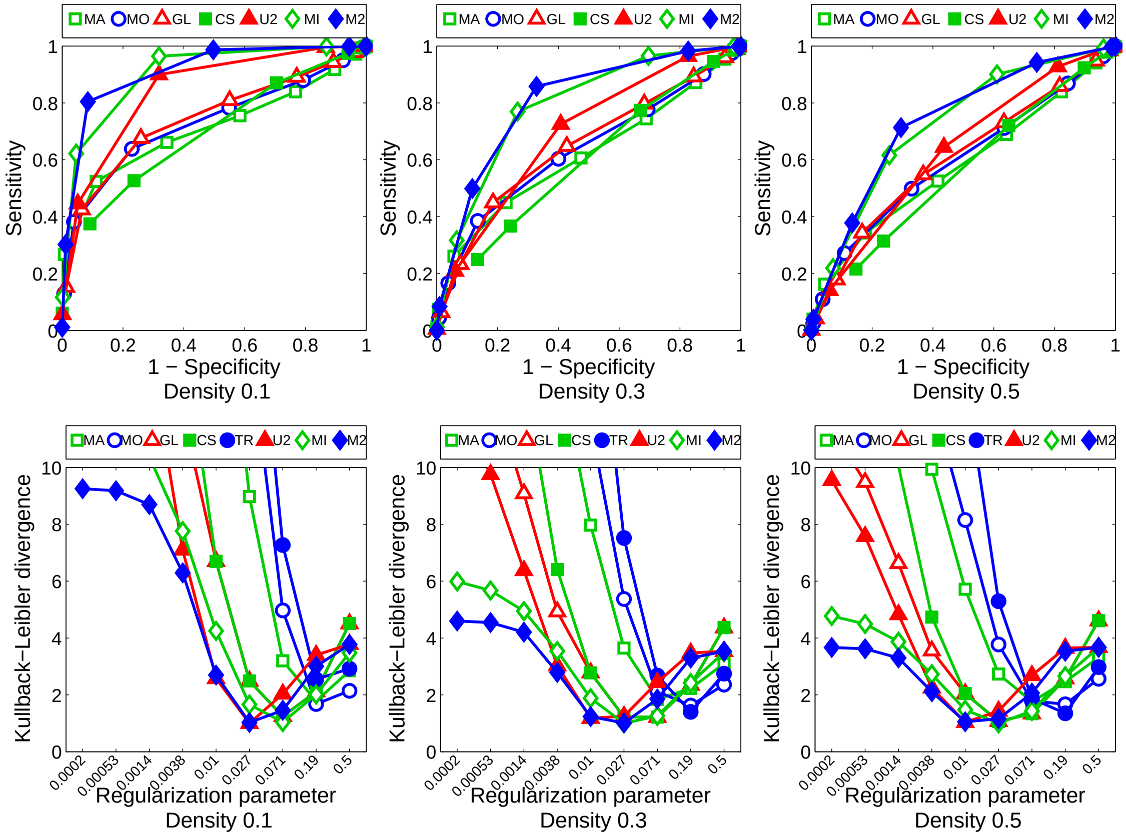


Figure 4.1: ROC curves (top) and cross-validated Kullback-Leibler divergence (bottom) between the recovered models and the ground truth for low (left), moderate (center) and high (right) edge density. Both our $\ell_{1,\infty}$ (MI) and $\ell_{1,2}$ (M2) multi-task methods recover the ground truth edges remarkably better than the $\ell_{1,2}$ multi-task upper bound method (U2), Meinshausen-Bühlmann with AND-rule (MA), OR-rule (MO), graphical lasso (GL), covariance selection (CS) and Tikhonov regularization (TR). The Kullback-Leibler divergence of our $\ell_{1,2}$ method is always lower than the $\ell_{1,2}$ upper bound method for all the regularization values.

upper bound method for all the regularization values. We did not observe a significant difference in Kullback-Leibler divergence between penalizing the weights in the diagonals versus not penalizing the diagonals for most regularization levels $\rho < 0.19$. For $\rho \geq 0.19$, diagonal penalization leads to a slightly worse Kullback-Leibler divergence. Furthermore, the ROC curves for our methods with and without diagonal penalization were the same. Therefore, we chose to report only the results without diagonal penalization.

In a second synthetic experiment, instead of producing the same topology $\boldsymbol{\Upsilon}_g$ for the models $\boldsymbol{\Omega}_g^{(k)}$ for all tasks k, we assume a level of similarity between the graph topologies. More specifically, a similarity of 1 means that the models $\boldsymbol{\Omega}_g^{(k)}$ have the same topology for all k, which is equivalent to our previous experiment. A similarity of 0 means that the models $\boldsymbol{\Omega}_g^{(k)}$ have different topology for all k. Figure 4.2 shows the ROC curves and Kullback-Leibler divergence between the recovered models and the ground truth. For a similarity of 0.5, both our $\ell_{1,\infty}$ and $\ell_{1,2}$ multi-task methods recover the ground truth edges as accurately as (similar ROC) the comparison methods. For higher similarity, our multi-task methods recover the

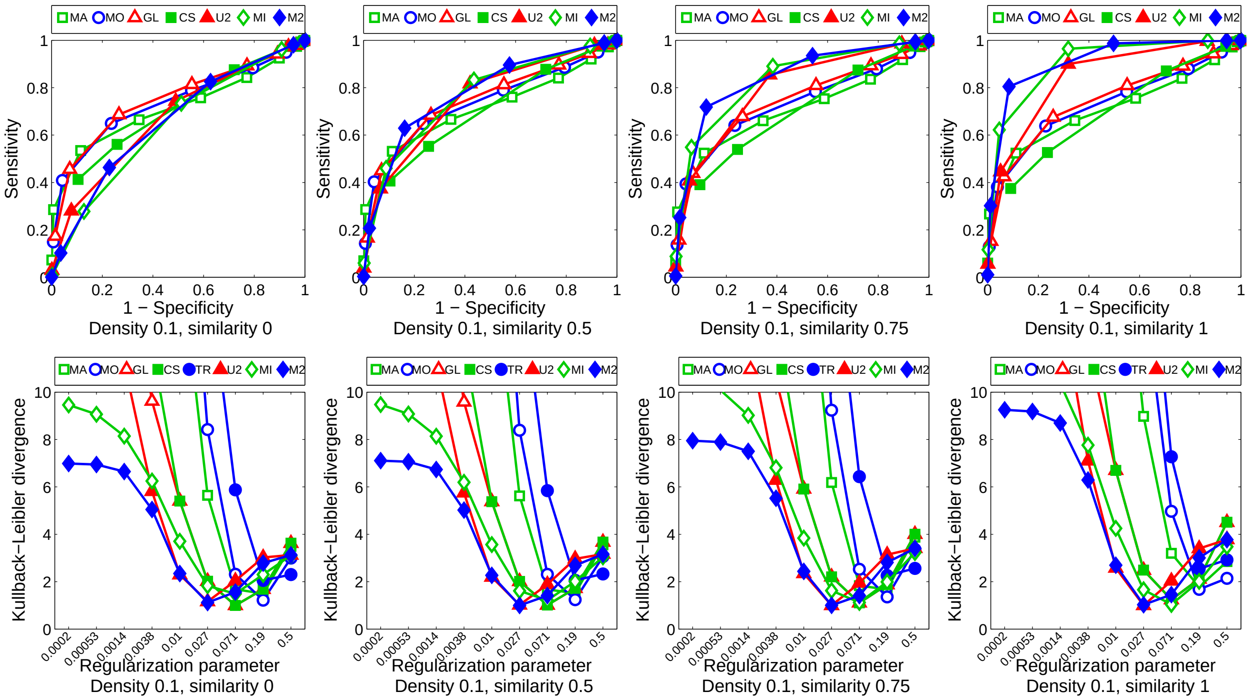
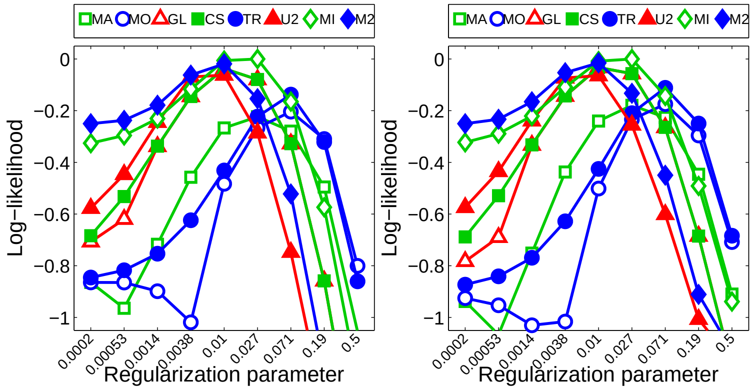


Figure 4.2: ROC curves (top) and cross-validated Kullback-Leibler divergence (bottom) between the recovered models and the ground truth for different levels of similarity across topologies. For a similarity of 0.5, both our $\ell_{1,\infty}$ (MI) and $\ell_{1,2}$ (M2) multi-task methods recover the ground truth edges as accurately as $\ell_{1,2}$ multi-task upper bound method (U2), Meinshausen-Bühlmann with AND-rule (MA), OR-rule (MO), graphical lasso (GL), covariance selection (CS) and Tikhonov regularization (TR). For higher similarity, our multi-task methods outperform the comparison methods. For lower similarity, the comparison methods outperform our multi-task methods. The Kullback-Leibler divergence of our $\ell_{1,2}$ method is always lower than the $\ell_{1,2}$ upper bound method for all the regularization values.

ground truth edges remarkably better (higher ROC) than the comparison methods. For lower similarity, the comparison methods outperform our multi-task methods. The latter behavior is expected given the small similarity of topologies in the ground truth models. As we will show later, real-world datasets seem to exhibit high similarity since our $\ell_{1,\infty}$ and $\ell_{1,2}$ multi-task methods outperform the comparison methods. Independently of the similarity of topologies, our $\ell_{1,2}$ method always produces better probability distributions (lower Kullback-Leibler divergence) than the $\ell_{1,2}$ upper bound method for all the regularization values.

For experimental validation on a real-world dataset, we first use a fMRI dataset that captures brain function of cocaine addicted and control subjects under conditions of monetary reward. The dataset collected by Goldstein et al. [2007] contains 16 cocaine addicted subjects and 12 control subjects. Six sessions were acquired for each subject. Each session contains 87 scans taken every 3.5 seconds. Registration of the dataset to the same spatial reference template (Talairach space) and spatial smoothing was performed in SPM2 (`http://www.fil.ion.ucl.ac.uk/spm/`). We extracted voxels from the gray matter only, and grouped them into 157 regions by using standard labels (Please, see Appendix E), given by the Talairach Daemon (`http://www.talairach.org/`). These regions span the entire



(a)

(b)

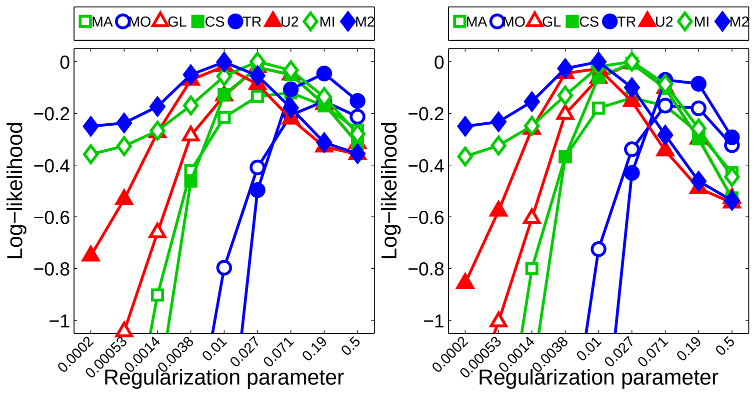
Figure 4.3: Cross-validated log-likelihood of structures learnt for each of the six sessions on cocaine addicted subjects (a) and control subjects (b). Both our $\ell_{1,\infty}$ (MI) and $\ell_{1,2}$ (M2) multi-task methods have higher log-likelihood than the $\ell_{1,2}$ multi-task upper bound method (U2), Meinshausen-Bühlmann with AND-rule (MA), OR-rule (MO), graphical lasso (GL), covariance selection (CS) and Tikhonov regularization (TR). Our $\ell_{1,2}$ method is always better than the $\ell_{1,2}$ multi-task upper bound method for all the regularization values.

brain (cerebellum, cerebrum and brainstem). In order to capture laterality effects, we have regions for the left and right side of the brain.

First, we test the idea of learning one Gaussian graphical model for each of the six sessions, i.e. each session is a task. We performed five-fold cross-validation on the subjects, and report the log-likelihood on the testing set (scaled for visualization purposes). In Figure 4.3, we can observe that the log-likelihood of both our $\ell_{1,\infty}$ and $\ell_{1,2}$ multi-task methods is better than the comparison methods. Moreover, our $\ell_{1,2}$ method is always better than the $\ell_{1,2}$ multi-task upper bound method of Varoquaux et al. [2010] for all the regularization values. We did not observe a significant difference in log-likelihood between penalizing the weights in the diagonals versus not penalizing the diagonals for most regularization levels $\rho < 0.19$. For $\rho \geq 0.19$, diagonal penalization leads to a slightly worse log-likelihood. Therefore, we chose to report only the results without diagonal penalization.

Second, we test the idea of learning one Gaussian graphical model for each subject, i.e. each subject is a task. It is well known that fMRI datasets have more variability across subjects than across sessions of the same subject. Therefore, our cross-validation setting works as follows: we use one session as training set, and the remaining five sessions as testing set. We repeat this procedure for all the six sessions and report the log-likelihood (scaled for visualization purposes). In Figure 4.4, we can observe that the log-likelihood of both our $\ell_{1,\infty}$ and $\ell_{1,2}$ multi-task methods is better than the comparison methods. Moreover, our $\ell_{1,2}$ method is always better than the $\ell_{1,2}$ multi-task upper bound method of Varoquaux et al. [2010] for all the regularization values. Finally, both our $\ell_{1,\infty}$ and $\ell_{1,2}$ methods are more stable for low regularization levels than the other methods in our evaluation, which perform very poorly.

In order to measure the statistical significance of our previously reported log-likelihoods, we further compared the best parameter setting for each of the techniques. In Tables 4.1 and 4.2, we report the two sample Z-statistic for the difference of both our $\ell_{1,\infty}$ and $\ell_{1,2}$ techniques



(a) (b)

Figure 4.4: Cross-validated log-likelihood of structures learnt for each subject on cocaine addicted subjects (a) and control subjects (b). Both our $\ell_{1,\infty}$ (MI) and $\ell_{1,2}$ (M2) multi-task methods have higher log-likelihood than the $\ell_{1,2}$ multi-task upper bound method (U2), Meinshausen-Bühlmann with AND-rule (MA), OR-rule (MO), graphical lasso (GL), covariance selection (CS) and Tikhonov regularization (TR). Our $\ell_{1,2}$ method is always better than the $\ell_{1,2}$ multi-task upper bound method for all the regularization values. For low regularization levels, our methods are more stable than the comparison methods.

Table 4.1: Z-statistic for the difference of log-likelihoods between our $\ell_{1,\infty}$ technique and each other method, for 16 cocaine addicted subjects. Except for few cases (marked with an asterisk), our method is statistically significantly better (90%, $Z > 1.28$) than the $\ell_{1,2}$ upper bound method (U2), Meinshausen-Bühlmann with AND-rule (MA), OR-rule (MO), graphical lasso (GL), covariance selection (CS) and Tikhonov regularization (TR). The $\ell_{1,\infty}$ and $\ell_{1,2}$ (M2) methods are not statistically significantly different.

Method	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**
MA	27.4	14.7	9.1	12.0	18.9	10.4	9.0	19.6	9.6	8.1	8.5	17.2	23.2	19.2	19.5	10.3
MO	25.6	17.0	10.4	13.7	19.4	10.4	10.7	20.4	13.5	11.2	10.1	18.5	22.5	17.6	21.1	13.9
GL	2.0	2.7	1.9	1.5	0.7*	1.8	2.7	2.2	4.3	2.9	1.8	1.8	1.8	2.2	4.7	2.9
CS	2.0	2.6	1.9	1.5	0.7*	1.8	2.7	2.1	4.3	2.9	1.8	1.8	1.8	2.2	4.7	2.9
TR	15.4	5.1	3.6	6.3	10.3	6.7	3.5	12.0	3.2	2.2	3.7	8.8	8.8	11.4	8.0	5.0
U2	5.5	2.3	1.7	2.0	4.0	1.9	1.1*	4.3	1.3	0.7*	1.3	3.9	3.3	3.9	2.9	1.2*
M2	0.8*	-0.3*	0.1*	0.3*	0.7*	0.5*	-0.1*	1.0*	-0.1*	-0.2*	-0.0*	0.8*	0.5*	1.1*	0.3*	-0.3*

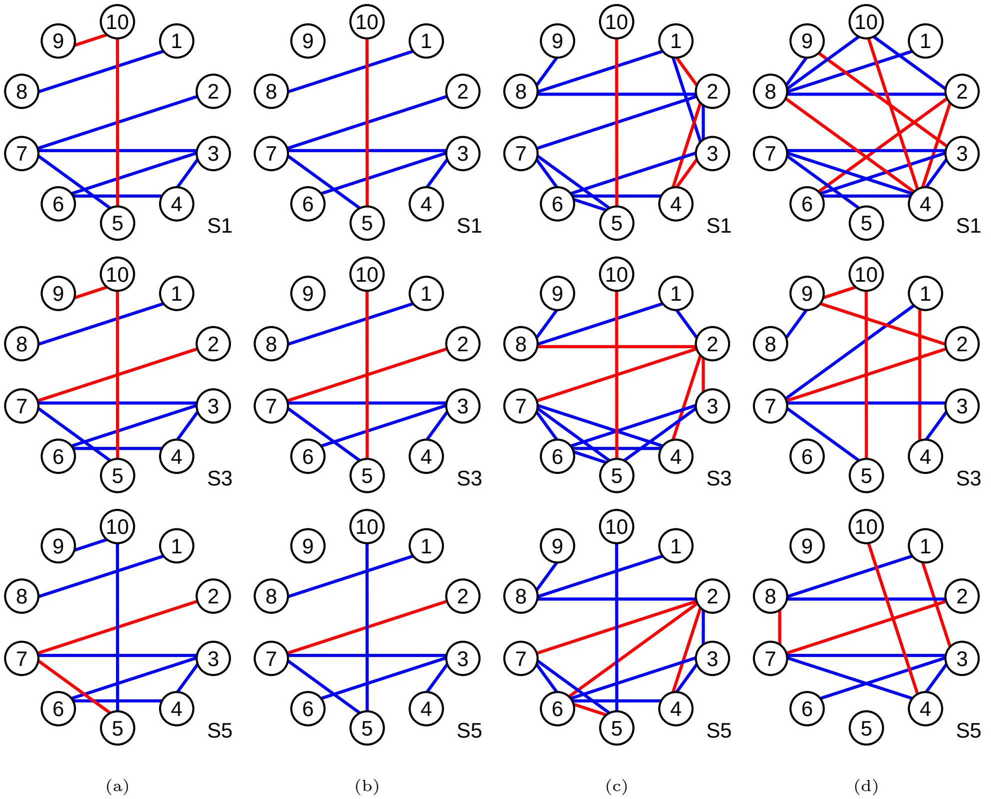
minus each competing method. Except for few subjects, the cross-validated log-likelihood of both our $\ell_{1,\infty}$ and $\ell_{1,2}$ methods is statistically significantly higher (90%, $Z > 1.28$) than the comparison methods, including the $\ell_{1,2}$ multi-task upper bound method of Varoquaux et al. [2010]. Our $\ell_{1,\infty}$ and $\ell_{1,2}$ multi-task methods are not statistically significantly different.

We show a subgraph of learnt structures for three randomly selected cocaine addicted subjects in Figure 4.5. We can observe that the sparseness pattern of the structures produced by our multi-task method is consistent across subjects.

Next, we present experimental results on a considerably larger real-world dataset. The *1000 functional connectomes* dataset contains resting-state fMRI of over 1128 subjects collected on several sites around the world. The dataset is publicly available at `http://www.nitrc.org/projects/fcon\_1000/`. Resting-state fMRI is a procedure that captures brain

Table 4.2: Z-statistic for the difference of log-likelihoods between our $\ell_{1,2}$ technique and each other method, for 16 cocaine addicted subjects. Except for few cases (marked with an asterisk), our method is statistically significantly better (90%, $Z > 1.28$) than the $\ell_{1,2}$ upper bound method (U2), Meinshausen-Bühlmann with AND-rule (MA), OR-rule (MO), graphical lasso (GL), covariance selection (CS) and Tikhonov regularization (TR). The $\ell_{1,\infty}$ (MI) and $\ell_{1,2}$ methods are not statistically significantly different.

Method	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**
MA	27.1	15.2	9.2	11.9	18.6	10.2	9.3	19.0	9.9	8.4	8.7	16.7	23.1	18.5	19.5	10.9
MO	25.3	17.5	10.5	13.7	19.1	10.2	10.9	19.8	13.8	11.5	10.3	18.1	22.4	16.9	21.1	14.4
GL	1.3	3.0	1.8	1.2*	0.0*	1.4	2.9	1.3*	4.5	3.1	1.9	1.1*	1.4	1.2*	4.5	3.3
CS	1.3	2.9	1.8	1.2*	0.0*	1.4	2.9	1.2*	4.5	3.1	1.9	1.0*	1.4	1.3*	4.5	3.2
TR	14.9	5.5	3.6	6.2	9.7	6.4	3.7	11.1	3.4	2.5	3.8	8.1	8.5	10.3	7.9	5.4
U2	4.8	2.6	1.7	1.8	3.3	1.5	1.2*	3.4	1.5	0.9*	1.4	3.2	2.8	2.9	2.6	1.5
MI	-0.8*	0.3*	-0.1*	-0.3*	-0.7*	-0.5*	0.1*	-1.0*	0.1*	0.2*	0.0*	-0.8*	-0.5*	-1.1*	-0.3*	0.3*



(a) (b) (c) (d)

Figure 4.5: Subgraph of ten randomly selected brain regions from learnt structures for three randomly selected cocaine addicted subjects, for (a) our $\ell_{1,\infty}$ multi-task method, (b) our $\ell_{1,2}$ multi-task method, (c) the $\ell_{1,2}$ upper bound method and (d) graphical lasso. Regularization parameter $\rho = 0.01$. Positive interactions are shown in blue, negative interactions are shown in red. Notice that sparseness of our structures (a,b) are consistent across subjects, while the remaining methods (c,d) fail to obtain a consistent sparseness pattern.

function of an individual that is not focused on the outside world, while his brain is at wakeful rest. Registration of the dataset to the same spatial reference template (Talairach space)

Table 4.3: Number of subjects per collection site and number of scans per subject in the *1000 functional connectomes* dataset.

Site	**Subjects**	**Scans**	**Site**	**Subjects**	**Scans**	**Site**	**Subjects**	**Scans**
AnnArbor_a	23	295	Cleveland1	17	125	NewYorkA2	24	192
Baltimore	46	120	Cleveland2	14	125	NewYorkB	20	168
Bangor	20	256	Dallas	24	114	Newark	19	135
Beijing1	40	225	ICBM	42	128	Ontario	11	100
Beijing2	42	225	Leiden1	12	210	Orangeburg	20	162
Beijing3	41	225	Leiden2	19	210	Oulu1	57	243
Beijing4	30	225	Leipzig	37	192	Oulu2	47	243
Beijing5	45	225	NYU_TRT1A	13	192	Oxford	22	175
Berlin	26	192	NYU_TRT1B	12	192	PaloAlto	17	234
Cambridge1	48	117	NYU_TRT2A	13	192	Queensland	18	189
Cambridge2	46	117	NYU_TRT2B	12	192	SaintLouis	31	125
Cambridge3	49	117	NYU_TRT3A	13	192	Taipei_a	13	256
Cambridge4	55	117	NYU_TRT3B	12	192	Taipei_b	8	160
CambridgeWG	35	144	NewYorkA1	35	192			

and spatial smoothing was performed in SPM2 (`http://www.fil.ion.ucl.ac.uk/spm/`). We extracted voxels from the gray matter only, and grouped them into 157 regions by using standard labels (Please, see Appendix E), given by the Talairach Daemon (`http://www.talairach.org/`). These regions span the entire brain (cerebellum, cerebrum and brainstem). In order to capture laterality effects, we have regions for the left and right side of the brain. Table 4.3 shows the number of subjects per collection site as well as the number of scans per subject.

We learn one Gaussian graphical model for each of the 41 collection sites, i.e. each site is a task. For each site, we used one third of the subjects for training, one third for validation and the remaining third for testing. We performed six repetitions by making each third of the subjects take turns as training, validation and testing sets. We report the negative log-likelihood on the testing set in Figure 4.6 (we subtracted the entropy measured on the testing set and then scaled the results for visualization purposes). We can observe that the log-likelihood of both our $\ell_{1,\infty}$ and $\ell_{1,2}$ multi-task methods is better than the comparison methods. Moreover, our $\ell_{1,2}$ method is better than the $\ell_{1,2}$ multi-task upper bound method of Varoquaux et al. [2010]. Our results also suggest that diagonal penalization does not produce better generalization performance.

Additionally in Figure 4.6, we also tested our previous regularizers: local constancy from Chapter 2 and $\ell_{1,2}$ variable selection from Chapter 3. We speculate that variable selection did not perform well because there seems to be a common structure across tasks (which is captured by our multi-task regularizer) and because the number of variables is relatively small (157 brain regions).

We show a subgraph of learnt structures for three randomly selected collection sites in Figure 4.7. We can observe that the sparseness pattern of the structures produced by our multi-task method is consistent across collection sites.

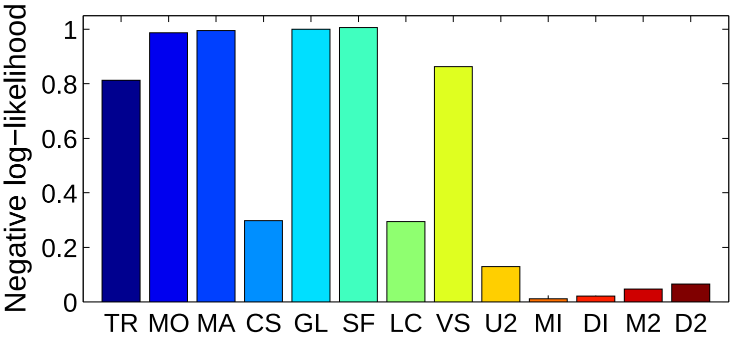


Figure 4.6: Test negative log-likelihood of structures learnt for the *1000 functional connectomes* dataset. Differences between our multi-task methods and the rest are statistically significant (99%, $Z > 2.33$). Both our $\ell_{1,\infty}$ (MI) and $\ell_{1,2}$ (M2) multi-task methods (without diagonal penalization) as well as our $\ell_{1,\infty}$ (DI) and $\ell_{1,2}$ (D2) multi-task methods (with diagonal penalization), have better log-likelihood than the $\ell_{1,2}$ multi-task upper bound method (U2), Meinshausen-Bühlmann with AND-rule (MA), OR-rule (MO), graphical lasso (GL), covariance selection (CS), Tikhonov regularization (TR) and scale free networks (SF). Our $\ell_{1,2}$ method is better than the $\ell_{1,2}$ multi-task upper bound method. Our results also suggest that diagonal penalization does not produce better generalization performance. Additionally, we also tested our previous regularizers: local constancy (LC) and $\ell_{1,2}$ variable selection (VS).

4.10 Concluding Remarks

In this chapter, we generalized the learning of sparse Gaussian graphical models to the multi-task setting by replacing the ℓ_1-norm regularization with an $\ell_{1,p}$-norm. We presented a block coordinate descent method which is provably convergent and yields sparse and positive definite estimates. We showed the connection between our $\ell_{1,\infty}$ multi-task structure learning problem and the continuous quadratic knapsack problem, as well as the connection between our $\ell_{1,2}$ multi-task structure learning problem and the quadratic trust-region problem. In synthetic experiments, we showed that our method outperforms others in recovering the topology of the ground truth model. The cross-validated log-likelihood of our method is higher than competing methods in two real-world brain fMRI datasets. For the $\ell_{1,2}$ problem, our block coordinate descent method leads to better ground-truth recovery and generalization when compared to the upper bound method of Varoquaux et al. [2010]. We experimentally found that diagonal penalization does not lead to a better generalization performance, when compared to not penalizing the diagonals. Our methods with and without diagonal penalization recover the ground truth edges similarly well. Therefore, we believe the negative impact of diagonal penalization is not on structure recovery but on parameter learning.

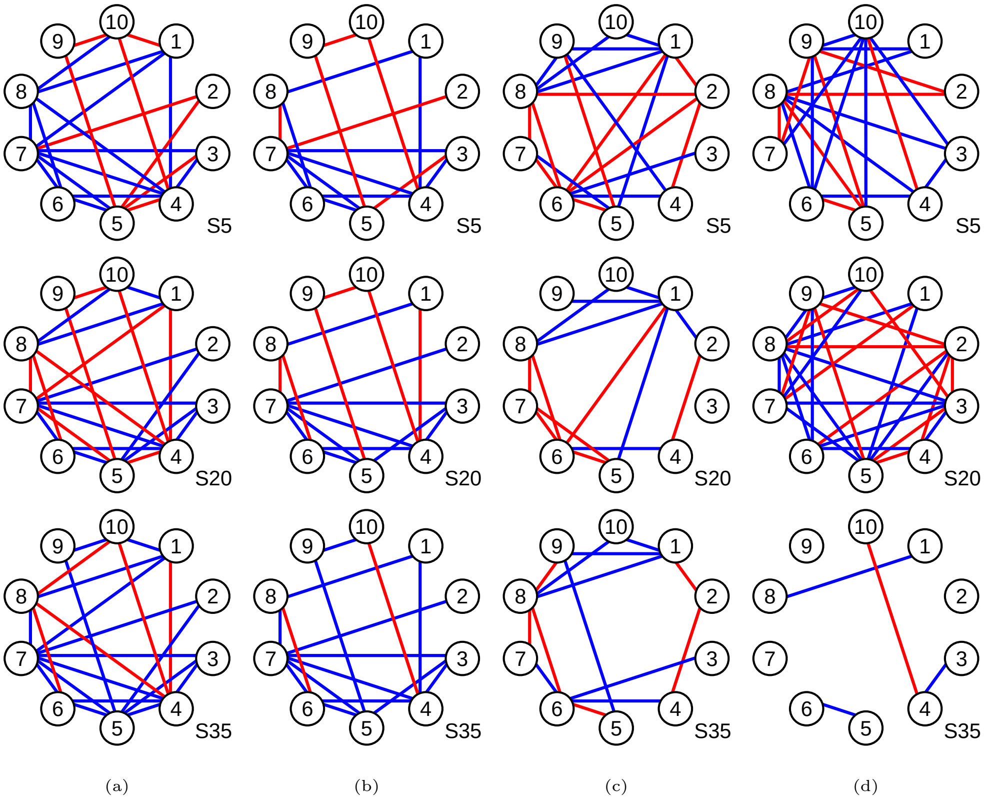


Figure 4.7: Subgraph of ten randomly selected brain regions from learnt structures for three randomly selected collection sites, for (a) our $\ell_{1,\infty}$ multi-task method, (b) our $\ell_{1,2}$ multi-task method, (c) the $\ell_{1,2}$ upper bound method and (d) covariance selection. Regularization parameter $\rho = 0.0002$. Positive interactions are shown in blue, negative interactions are shown in red. Notice that sparseness of our structures (a,b) are consistent across collection sites, while the remaining methods (c,d) fail to obtain a consistent sparseness pattern.

Chapter 5

Learning Discrete MRFs

In the previous chapters, we focused on learning structures for continuous and jointly Gaussian variables. In this chapter, we focus on learning structures for discrete variables. In fact, our results are far more general than the specific problem of structure learning. Our main contribution is in the area of optimization, since we analyze the problem of *biased stochastic optimization.*

We study the convergence rate of *stochastic* optimization of *exact* (NP-hard) objectives, for which only biased estimates of the gradient are available. We motivate this problem in the context of learning the structure and parameters of Ising models. We first provide a convergence-rate analysis of *deterministic* errors for *forward-backward splitting* (FBS). We then extend our analysis to *biased stochastic* errors, by first characterizing a family of samplers and providing a high probability bound that allows understanding not only FBS, but also *proximal gradient* (PG) methods. We derive some interesting conclusions: FBS requires only a logarithmically increasing number of random samples in order to converge (although at a very low rate); the required number of random samples is the same for the deterministic and the biased stochastic setting for FBS and basic PG; accelerated PG is not guaranteed to converge in the biased stochastic setting.

5.1 Introduction

One challenge of structure learning is that the number of possible structures is super-exponential in the number of variables. For Ising models, the number of parameters, the number of edges in the structure and the number of non-zero elements in the *ferro-magnetic coupling* matrix are equivalent measures of model complexity. Therefore a computationally tractable approach is to use sparseness promoting regularizers [Wainwright et al., 2006, Banerjee et al., 2008, Höfling and Tibshirani, 2009].

One additional challenge for Ising models (and Markov random fields in general) is that computing the likelihood of a candidate structure is NP-hard. For this reason, several researchers propose exact optimization of approximate objectives, such as ℓ_1-regularized logistic regression [Wainwright et al., 2006], greedy optimization of the conditional log-likelihoods [Jalali et al., 2011], pseudo-likelihood [Besag, 1975] and a sequence of first-order approximations of the exact log-likelihood [Höfling and Tibshirani, 2009]. Several convex upper bounds and approximations to the log-partition function have been proposed for maximum likelihood estimation, such as the log-determinant relaxation [Banerjee et al., 2008], the cardinality bound [El Ghaoui and Gueye, 2008], the Bethe entropy [Lee et al., 2006a, Parise and Welling, 2006], tree-reweighted approximations and general weighted free-energy [Yang and Ravikumar, 2011].

In this chapter, we focus on the stochastic optimization of the exact log-likelihood as our motivating problem. The use of *stochastic maximum likelihood* dates back to [Geyer, 1991, Younes, 1988], in which Markov chain Monte Carlo (MCMC) was used for approximating the gradient. For restricted Boltzmann machines (a very related graphical model) researchers have proposed a variety of approximation methods, such as variational approximations [Murray and Ghahramani, 2004], contrastive divergence [Hinton, 2002], persistent contrastive divergence [Tieleman, 2008], tempered MCMC [Salakhutdinov, 2009, Desjardins et al., 2010], adaptive MCMC [Salakhutdinov, 2010] and particle filtering [Asuncion et al., 2010].

Empirical results in [Marlin et al., 2010] suggests that stochastic maximum likelihood is superior to contrastive divergence, pseudo-likelihood, ratio matching and generalized score matching for learning restricted Boltzmann machines, in the sense that it produces a higher test set log-likelihood, and more consistent classification results across datasets.

Learning sparse Ising models leads to the use of stochastic optimization with biased estimates of the gradient. Most work in stochastic optimization assumes the availability of unbiased estimates [Duchi and Singer, 2009c, Duchi et al., 2010, Hu et al., 2009, Nemirovski et al., 2009, Duchi and Singer, 2009c, Duchi et al., 2010, Hu et al., 2009, Nemirovski et al., 2009, Langford et al., 2009, Shalev-Shwartz et al., 2007, Shalev-Shwartz and Tewari, 2009] Additionally, other researchers have analyzed convergence rates in the presence of *deterministic* errors that do not decrease over time [d'Aspremont, 2008, Baes, 2009, Devolder et al., 2011] and show convergence up to a constant level. Similarly, Devolder [2012] analyzed the case of *stochastic* errors with *fixed* bias and variance and show convergence up to a constant level.

Notable exceptions are the recent works of Schmidt et al. [2011], Friedlander and Schmidt [2011], Duchi et al. [2011]. Schmidt et al. [2011] analyzed *proximal-gradient* (PG) methods for *deterministic* errors of the gradient that decrease over time, for inexact projection steps and Lipschitz as well as strongly convex functions. In our work, we restrict our analysis to exact projection steps and do not assume strong convexity. Both assumptions are natural for learning sparse models under the ℓ_1 regularization. Friedlander and Schmidt [2011] provides convergence rates in expected value for PG with *stochastic* errors that decrease over time in expected value. Friedlander and Schmidt [2011] proposes a growing sample-size strategy for approximating the gradient, i.e. by picking an increasing number of training samples in order to better approximate the gradient. In contrast, our work for is for NP-hard gradients and we provide bounds with high probability, by taking into account the bias and the variance of the errors. Duchi et al. [2011] analyzed *mirror descent* (a generalization that includes forward-backward splitting) and show convergence rates in expected value and with high probability with respect to the mixing time of the sampling distribution. We argue that practitioners usually terminate Markov chains before properly mixing, and therefore we motivate our analysis for a controlled increasing number of random samples.

Regarding our contribution in optimization, we provide a convergence-rate analysis of *deterministic* errors for three different flavors of *forward-backward splitting* (FBS): robust [Nemirovski et al., 2009], basic and random [Duchi and Singer, 2009c]. We extend our analysis to *biased stochastic* errors, by first characterizing a family of samplers (including importance sampling and MCMC) and providing a high probability bound that is useful for understanding the convergence of not only FBS, but also PG [Schmidt et al., 2011]. Our analysis shows the bias/variance term and allow to derive some interesting conclusions. First, FBS for deterministic or biased stochastic errors requires only a logarithmically increasing number of random samples in order to converge (although at a very low rate). More interestingly, we found that the required number of random samples is the same for the deterministic and the biased stochastic setting for FBS and basic PG. We also found that accelerated PG is not guaranteed to converge in the biased stochastic setting.

Regarding our contribution in structure learning, we show that the optimal solution of maximum likelihood estimation is bounded (to the best of our knowledge, this has not been shown before). Our analysis shows provable convergence guarantees for finite iterations and finite number of random samples. Note that while consistency in structure recovery has been established (e.g. Wainwright et al. [2006]), convergence rates of parameter learning for fixed structures is up to now unknown. Our analysis can be easily extended to Markov random fields with higher order cliques as well as parameter learning for fixed structures by using a ℓ_2^2 regularizer instead.

Last but not least, it is important to mention that in the context of learning Markov random fields, other techniques have been developed. Note that we are not developing a new technique and our goal is not to prove the superiority of neither of the existing techniques, but to analyze the convergence properties. That being said, another approach to learn Markov random fields consist in iteratively performing conditional independence tests between pairs of variables. Several heuristics have been proposed for the order of independence tests: *edge deletion method* and the *Markov blanket method* in [Pearl, 1988] as well as the *grow-shrink method* of [Bromberg et al., 2006]. One additional approach is the algorithm of Roy et al. [2009] that learns the structure by finding the best neighborhood or Markov blanket for each node, and that relies in the canonical parametrization of factor graphs [Abbeel et al., 2005]. Finally, a *maximum entropy relaxation* approach was proposed in [Johnson et al., 2007].

5.2 Our Motivating Problem

In this section, we introduce the problem of learning sparse Ising models and discuss its properties. Our discussion will motivate a set of bounds and assumptions for a more general convergence rate analysis.

5.2.1 Problem Setup

An *Ising model* is a Markov random field on binary variables with pairwise interactions. It first arose in statistical physics as a model for the energy of a physical system of interacting atoms [Ising, 1925, Koller and Friedman, 2009]. Formally, the probability mass function

(PMF) of an Ising model parameterized by $\boldsymbol{\theta} = (\mathbf{W}, \mathbf{b})$ is defined as:

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = \frac{1}{\mathcal{Z}(\mathbf{W}, \mathbf{b})} e^{\mathbf{x}^{\mathrm{T}}\mathbf{W}\mathbf{x} + \mathbf{b}^{\mathrm{T}}\mathbf{x}} \tag{5.1}$$

where the domain for the binary variables is $\mathbf{x} \in \{-1, +1\}^N$, $\mathbf{W} \in \mathbb{R}^{N \times N}$ is symmetric with zero diagonal (i.e. $\mathbf{diag}(\mathbf{W}) = \mathbf{0}$), $\mathbf{b} \in \mathbb{R}^N$ and partition function is defined as $\mathcal{Z}(\mathbf{W}, \mathbf{b}) = \sum_{\mathbf{x}} e^{\mathbf{x}^{\mathrm{T}}\mathbf{W}\mathbf{x} + \mathbf{b}^{\mathrm{T}}\mathbf{x}}$. For clarity of the convergence rate analysis, we also define $\boldsymbol{\theta} \in \mathbb{R}^M$ where $M = N^2$.

In the physics literature, $\mathbf{W}$ and $\mathbf{b}$ are called *ferro-magnetic coupling* and *external magnetic field* respectively [Ising, 1925]. $\mathbf{W}$ defines the topology of the Markov random field, i.e. the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is defined as $\mathcal{V} = \{1, \ldots, N\}$ and $\mathcal{E} = \{(n_1, n_2) \mid n_1 < n_2 \wedge w_{n_1 n_2} \neq 0\}$. It is well known that, for an Ising model with arbitrary topology, computing the partition function $\mathcal{Z}$ is NP-hard [Barahona, 1982]. It is also NP-hard to approximate $\mathcal{Z}$ with high probability and arbitrary precision [Chandrasekaran et al., 2008].

Conditional independence in an Ising model is simply reflected in the zero entries of the *ferro-magnetic coupling* matrix $\mathbf{W}$, i.e. two variables n_1 and n_2 are conditionally independent if and only if $w_{n_1 n_2} = 0$. The number of edges $|\mathcal{E}|$ or equivalently the cardinality (number of non-zero entries) of $\mathbf{W}$ is a measure of model complexity, and it can be used as a regularizer for maximum likelihood estimation. The main disadvantage of using such penalty is that it leads to a NP-hard problem, regardless of the computational complexity of the log-likelihood.

Next, we formalize the problem of finding a sparse Ising model by regularized maximum likelihood estimation. We replace the cardinality penalty by the ℓ_1-norm regularizer as in [Wainwright et al., 2006, Banerjee et al., 2008, Höfling and Tibshirani, 2009].

Given a complete dataset with T i.i.d. samples $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)}$, and a sparseness parameter $\rho > 0$ the ℓ_1-regularized maximum likelihood estimation for the Ising model in eq.(5.1) becomes:

$$\min_{\mathbf{W}, \mathbf{b}} \mathcal{L}(\mathbf{W}, \mathbf{b}) + \mathcal{R}(\mathbf{W}) \tag{5.2}$$

where the negative (average) log-likelihood $\mathcal{L}(\mathbf{W}, \mathbf{b}) = -\frac{1}{T}\sum_t \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(t)}) = \log \mathcal{Z}(\mathbf{W}, \mathbf{b}) - \langle \widehat{\boldsymbol{\Sigma}}, \mathbf{W} \rangle - \widehat{\boldsymbol{\mu}}^{\mathrm{T}}\mathbf{b}$, the empirical second-order moment $\widehat{\boldsymbol{\Sigma}} = \frac{1}{T}\sum_t \mathbf{x}^{(t)}\mathbf{x}^{(t)\mathrm{T}} - \mathbf{I}$, the empirical first-order moment $\widehat{\boldsymbol{\mu}} = \frac{1}{T}\sum_t \mathbf{x}^{(t)}$ and the regularizer $\mathcal{R}(\mathbf{W}) = \rho \|\mathbf{W}\|_1$.

The objective function in eq.(5.2) is convex, given the convexity of the log-partition function [Koller and Friedman, 2009], linearity of the scalar products and convexity of the non-smooth ℓ_1-norm regularizer. As discussed before, computing the partition function $\mathcal{Z}$ is NP-hard, and so is computing the objective function in eq.(5.2).

5.2.2 Bounds

In what follows, we show boundedness of the optimal solution and the gradients of the maximum likelihood problem. Both are important ingredients for showing convergence and are largely used assumptions in optimization. In this chapter, we follow the original formulation of the problem given in [Wainwright et al., 2006, Banerjee et al., 2008, Höfling and Tibshirani, 2009], which does not regularize $\mathbf{b}$. We found interesting to show that this problem has bounds for $\|\mathbf{b}^*\|_1$ unlike other stochastic optimization problems, e.g. SVMs [Shalev-Shwartz et al., 2007].

First, we make some observations that will help us derive our bounds. The empirical second-order moment $\widehat{\mathbf{\Sigma}}$ and first-order moment $\widehat{\boldsymbol{\mu}}$ in eq.(5.2) are computed from binary variables in $\{-1, +1\}$, therefore $\|\widehat{\mathbf{\Sigma}}\|_\infty \leq 1$ and $\|\widehat{\boldsymbol{\mu}}\|_\infty \leq 1$.

Assumption 5.1. *It is reasonable to assume that the empirical first-order moment of every variable is not equal to* -1 *(or* $+1$*), since this would be equivalent to observe a constant value* -1 *(or* $+1$*) for such variables in every sample in the dataset, i.e.* $(\exists n)\ |\widehat{\mu}_n| = 1 \Leftrightarrow (\forall t)\ x_n^{(t)} = -1 \vee (\forall t)\ x_n^{(t)} = 1$. *Therefore, we assume* $\|\widehat{\boldsymbol{\mu}}\|_\infty < 1 \Leftrightarrow (\forall n) - 1 < \widehat{\mu}_n < +1$.

Given those observations, we state our bounds in the following theorem. For clarity of the convergence rate analysis, we also define the bound D of the optimal solution.

Theorem 5.2. *The optimal solution* $\boldsymbol{\theta}^* = (\mathbf{W}^*, \mathbf{b}^*)$ *of the maximum likelihood problem in eq.*(5.2) *is bounded as follows:*

$$\begin{array}{lll} \text{i.} & \|\mathbf{W}^*\|_1 & \leq \frac{N \log 2}{\rho} \\ \text{ii.} & \|\mathbf{b}^*\|_1 & \leq \frac{N \log 2(\rho+1+\|\widehat{\mathbf{\Sigma}}\|_\infty)}{\rho(1-\|\widehat{\boldsymbol{\mu}}\|_\infty)} \\ \text{iii.} & \|\boldsymbol{\theta}^*\|_2 & \leq D \end{array} \tag{5.3}$$

where $D^2 = \left(\frac{N \log 2}{\rho}\right)^2 \left(1 + \left(\frac{\rho+1+\|\widehat{\mathbf{\Sigma}}\|_\infty}{1-\|\widehat{\boldsymbol{\mu}}\|_\infty}\right)^2\right)$.

Proof. For proving Claim i, note that for Ising models (and in general for any discrete probability distribution) the negative log-likelihood in eq.(5.2) is non-negative, i.e. $(\forall \mathbf{x})\ p_{\boldsymbol{\theta}}(\mathbf{x}) \in [0; 1] \Rightarrow (\forall \mathbf{x})\ \log p_{\boldsymbol{\theta}}(\mathbf{x}) \leq 0 \Rightarrow \mathcal{L}(\mathbf{W}, \mathbf{b}) = -\frac{1}{T}\sum_t \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(t)}) \geq 0$. Given that $(\mathbf{W}^*, \mathbf{b}^*)$ is the optimal solution, $N \log 2 = \mathcal{L}(\mathbf{0}, \mathbf{0}) + \mathcal{R}(\mathbf{0}) \geq \mathcal{L}(\mathbf{W}^*, \mathbf{b}^*) + \mathcal{R}(\mathbf{W}^*) \geq \mathcal{R}(\mathbf{W}^*) = \rho\|\mathbf{W}^*\|_1$, and we prove our claim.

For proving Claim ii, note that the regularizer $\mathcal{R}(\mathbf{W})$ is non-negative, therefore $N \log 2 = \mathcal{L}(\mathbf{0}, \mathbf{0}) + \mathcal{R}(\mathbf{0}) \geq \mathcal{L}(\mathbf{W}^*, \mathbf{b}^*) + \mathcal{R}(\mathbf{W}^*) \geq \mathcal{L}(\mathbf{W}^*, \mathbf{b}^*) \geq \log(\sum_{\mathbf{x}} e^{-\|\mathbf{W}^*\|_1 + \mathbf{b}^{*\mathrm{T}}\mathbf{x}}) - \|\widehat{\mathbf{\Sigma}}\|_\infty \|\mathbf{W}^*\|_1 - \widehat{\boldsymbol{\mu}}^{\mathrm{T}}\mathbf{b}^* = -\|\mathbf{W}^*\|_1 + \log(\sum_{\mathbf{x}} e^{\mathbf{b}^{*\mathrm{T}}\mathbf{x}}) - \|\widehat{\mathbf{\Sigma}}\|_\infty \|\mathbf{W}^*\|_1 - \widehat{\boldsymbol{\mu}}^{\mathrm{T}}\mathbf{b}^* = \sum_n \log(e^{b_n^*} + e^{-b_n^*}) - \widehat{\boldsymbol{\mu}}^{\mathrm{T}}\mathbf{b}^* - (1 + \|\widehat{\mathbf{\Sigma}}\|_\infty)\|\mathbf{W}^*\|_1 \geq \|\mathbf{b}^*\|_1 - \widehat{\boldsymbol{\mu}}^{\mathrm{T}}\mathbf{b}^* - (1 + \|\widehat{\mathbf{\Sigma}}\|_\infty)\|\mathbf{W}^*\|_1 \geq (1 - \|\widehat{\boldsymbol{\mu}}\|_\infty)\|\mathbf{b}^*\|_1 - (1 + \|\widehat{\mathbf{\Sigma}}\|_\infty)\|\mathbf{W}^*\|_1$. Recall that by Assumption 5.1, $\|\widehat{\boldsymbol{\mu}}\|_\infty < 1$. Therefore, $\|\mathbf{b}^*\|_1 \leq (N \log 2 + (1 + \|\widehat{\mathbf{\Sigma}}\|_\infty)\|\mathbf{W}^*\|_1)/(1 - \|\widehat{\boldsymbol{\mu}}\|_\infty)$ and by using Claim i we prove our claim.

Claim iii follows from Claims i and ii and the fact that $\|\boldsymbol{\theta}^*\|_2^2 = \|\mathbf{W}^*\|_{\mathfrak{F}}^2 + \|\mathbf{b}^*\|_2^2 \leq \|\mathbf{W}^*\|_1^2 + \|\mathbf{b}^*\|_1^2$. □

If we choose to add the regularizer $\rho\|\mathbf{b}\|_1$ in eq.(5.2), it is easy to conclude that $\|\mathbf{W}^*\|_1 + \|\mathbf{b}^*\|_1 \leq \frac{N \log 2}{\rho}$ as in Claim i of Theorem 5.2.

The gradient of the objective function of the maximum likelihood problem in eq.(5.2) is defined as:

$$\begin{array}{lll} \text{i.} & \partial \log \mathcal{Z}/\partial \mathbf{W} & = \mathbb{E}_{\mathcal{P}}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] \\ \text{ii.} & \partial \log \mathcal{Z}/\partial \mathbf{b} & = \mathbb{E}_{\mathcal{P}}[\mathbf{x}] \\ \text{iii.} & \partial \mathcal{L}/\partial \mathbf{W} & = \partial \log \mathcal{Z}/\partial \mathbf{W} - \widehat{\mathbf{\Sigma}} \\ \text{iv.} & \partial \mathcal{L}/\partial \mathbf{b} & = \partial \log \mathcal{Z}/\partial \mathbf{b} - \widehat{\boldsymbol{\mu}} \end{array} \tag{5.4}$$

where $\mathcal{P}$ is the probability distribution with PMF $p_{\boldsymbol{\theta}}(\mathbf{x})$. The expression in eq.(5.4) uses the fact that $\mathbb{E}_{\mathcal{P}}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] = \sum_{\mathbf{x}} \mathbf{x}\mathbf{x}^{\mathrm{T}} p_{\boldsymbol{\theta}}(\mathbf{x})$ and $\mathbb{E}_{\mathcal{P}}[\mathbf{x}] = \sum_{\mathbf{x}} \mathbf{x} p_{\boldsymbol{\theta}}(\mathbf{x})$.

It is well known that computing the gradients $\partial \log \mathcal{Z}/\partial \mathbf{W}$ and $\partial \log \mathcal{Z}/\partial \mathbf{b}$ is NP-hard. The complexity results in [Chandrasekaran et al., 2008] imply that approximating those gradients with high probability and arbitrary precision is also NP-hard.

Next, we state some properties of the gradient of the exact log-likelihood. For clarity of the convergence rate analysis, we also define the Lipschitz constant G.

Lemma 5.3. *The objective function of the maximum likelihood problem in eq.*(5.2) *has the following Lipschitz continuity properties:*

$$\begin{array}{ll} \text{i.} & \|\partial \log \mathcal{Z}/\partial \mathbf{W}\|_\infty \,,\ \|\partial \log \mathcal{Z}/\partial \mathbf{b}\|_\infty \leq 1 \\ \text{ii.} & \|\partial \mathcal{L}/\partial \mathbf{W}\|_\infty \ \leq 1 + \|\widehat{\mathbf{\Sigma}}\|_\infty \\ \text{iii.} & \|\partial \mathcal{L}/\partial \mathbf{b}\|_\infty \ \ \leq 1 + \|\widehat{\boldsymbol{\mu}}\|_\infty \\ \text{iv.} & \|\partial \mathcal{R}/\partial \mathbf{W}\|_\infty \leq \rho \\ \text{v.} & \|\partial \mathcal{L}/\partial \boldsymbol{\theta}\|_2 \,,\ \|\partial \mathcal{R}/\partial \boldsymbol{\theta}\|_2 \leq G \end{array} \tag{5.5}$$

where $G^2 = \max(N^2(1 + \|\widehat{\mathbf{\Sigma}}\|_\infty)^2 + N(1 + \|\widehat{\boldsymbol{\mu}}\|_\infty)^2, N^2\rho^2)$.

Proof. For proving Claim i, note that the terms $\partial \log \mathcal{Z}/\partial \mathbf{W}$ and $\partial \log \mathcal{Z}/\partial \mathbf{b}$ in eq.(5.4) are the second and first-order moment of binary variables in $\{-1, +1\}$.

Proving Claims ii and iii is straightforward from applying the above claims in eq.(5.4).

For proving Claim iv, recall that the subgradient $\partial \mathcal{R}/\partial \mathbf{W} = \{\mathbf{G} \mid \|\mathbf{G}\|_\infty \leq \rho \wedge \langle \mathbf{G}, \mathbf{W}\rangle = \|\mathbf{W}\|_1\}$. Therefore, $(\forall \mathbf{G} \in \partial \mathcal{R}/\partial \mathbf{W})\ \|\mathbf{G}\|_\infty \leq \rho$.

Claim v follows from Claims ii to iv and the fact that $\|\partial \mathcal{L}/\partial \boldsymbol{\theta}\|_2^2 = \|\partial \mathcal{L}/\partial \mathbf{W}\|_{\mathfrak{F}}^2 + \|\partial \mathcal{L}/\partial \mathbf{b}\|_2^2$. Furthermore, for the first term $\|\partial \mathcal{L}/\partial \mathbf{W}\|_{\mathfrak{F}} \leq N\|\partial \mathcal{L}/\partial \mathbf{W}\|_\infty \leq N(1 + \|\widehat{\mathbf{\Sigma}}\|_\infty)$, and for the second term $\|\partial \mathcal{L}/\partial \mathbf{b}\|_2 \leq \sqrt{N}\|\partial \mathcal{L}/\partial \mathbf{b}\|_\infty \leq \sqrt{N}(1+\|\widehat{\boldsymbol{\mu}}\|_\infty)$. Similarly, $\|\partial \mathcal{R}/\partial \boldsymbol{\theta}\|_2 = \|\partial \mathcal{R}/\partial \mathbf{W}\|_{\mathfrak{F}} \leq N\|\partial \mathcal{R}/\partial \mathbf{W}\|_\infty \leq N\rho$. □

5.2.3 Approximating the Gradient of the Log-Partition Function

Suppose one wants to evaluate the expression $\mathbb{E}_{\mathcal{P}}[\mathbf{x}\mathbf{x}^{\mathrm{T}}]$ in eq.(5.4) which is the gradient of the log-partition function. Let assume we know the distribution $p_{\boldsymbol{\theta}}(\mathbf{x})$ up to a constant factor, i.e. $p'_{\boldsymbol{\theta}}(\mathbf{x}) = e^{\mathbf{x}^{\mathrm{T}}\mathbf{W}\mathbf{x}+\mathbf{b}^{\mathrm{T}}\mathbf{x}}$. Importance sampling draws S samples $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(S)}$ from a trial distribution with PMF $q(\mathbf{x})$, calculates the importance weights $\alpha^{(s)} = p'_{\boldsymbol{\theta}}(\mathbf{x}^{(s)})/q(\mathbf{x}^{(s)})$ and produces the estimate $(\sum_s \alpha^{(s)}\mathbf{x}^{(s)}\mathbf{x}^{(s)\mathrm{T}})/\sum_s \alpha^{(s)}$. On the other hand, MCMC generates S samples $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(S)}$ from the distribution $p_{\boldsymbol{\theta}}(\mathbf{x})$ based on constructing a Markov chain whose stationary distribution is $p_{\boldsymbol{\theta}}(\mathbf{x})$. Thus, the estimate becomes $\frac{1}{S}\sum_s \mathbf{x}^{(s)}\mathbf{x}^{(s)\mathrm{T}}$.

In what follows, we characterize a family of samplers that includes importance sampling and MCMC as shown in [Peskun, 1973, Liu, 2001].

Definition 5.4. *A* (B, V, S, D)*-sampler takes* S *random samples from a distribution* $\mathcal{Q}$ *and produces biased estimates of the gradients of the log-partition function* $\partial \log \mathcal{Z}/\partial \boldsymbol{\theta} + \boldsymbol{\xi}$*, with error* $\boldsymbol{\xi}$ *that has bias and variance:*

$$\begin{array}{ll} \text{i.} & \mathbb{E}_{\mathcal{Q}}[\|\boldsymbol{\xi}\|_2] \quad \leq \frac{B}{S} + \mathcal{O}(\frac{1}{S^2}) \\ \text{ii.} & \mathbb{V}\mathrm{ar}_{\mathcal{Q}}[\|\boldsymbol{\xi}\|_2] \leq \frac{V}{S} + \mathcal{O}(\frac{1}{S^2}) \end{array} \tag{5.6}$$

for $B \geq 0, V \geq 0$ *and* $(\forall \boldsymbol{\theta})\ \|\boldsymbol{\theta}\|_2 \leq D$.

Note that a (B, V, S, D)-sampler is asymptotically unbiased with asymptotically vanishing variance, i.e. $S \to +\infty \Rightarrow \frac{B}{S} \to 0 \wedge \frac{V}{S} \to 0$. Unfortunately, analytical approximations of the constants B and V are difficult to obtain even for specific classes, e.g. Ising models. The theoretical analysis implies that such constants B and V exist [Peskun, 1973, Liu, 2001] for importance sampling and MCMC. We argue that this apparent disadvantage does not diminish the relevance of our analysis, since we can reasonably expect that more refined samplers lead to lower B and V.

Note that Definition 5.4 does not contradict the complexity results in [Chandrasekaran et al., 2008] that show that it is likely impossible to approximate $\mathcal{Z}$ (and therefore its gradient) with probability greater than $1 - \delta$ and arbitrary precision ε in time polynomial in $\log \frac{1}{\delta}$ and $\frac{1}{\varepsilon}$. Definition 5.4 assumes biasedness and a polynomial decay instead of an exponential decay (which is a more stringent condition) and cannot be used to derive two-sided high probability bounds that are both $\mathcal{O}(\log \frac{1}{\delta})$ and $\mathcal{O}(\frac{1}{S})$. Therefore, Definition 5.4 cannot be used to obtain polynomial-time algorithms as the ones considered in [Chandrasekaran et al., 2008].

Assumption 5.5. *It is reasonable to assume that the estimates of the gradient of the log-partition function are inside* $[-1; +1]$ *since they are approximations of the second and first-order moment of binary variables in* $\{-1, +1\}$. *Furthermore, it is straightforward to enforce Lipschitz continuity (condition i of Lemma 5.3) for any sampler (e.g. importance sampling, MCMC or any conceivable method) by limiting its output to be inside* $[-1; +1]$. *More formally, we have:*

$$\begin{array}{lll} \text{i.} & \|\partial \log \mathcal{Z}/\partial \boldsymbol{\theta} + \boldsymbol{\xi}\|_\infty & \leq 1 \\ \text{ii.} & \|\partial \mathcal{L}/\partial \boldsymbol{\theta} + \boldsymbol{\xi}\|_2 & \leq G \end{array} \tag{5.7}$$

5.3 Biased Stochastic Optimization

In this section, we analyze the convergence rates of *forward-backward splitting.* Our results apply to any problem that fulfills the following largely used assumptions in optimization:

- the objective function is composed by a smooth function $\mathcal{L}(\boldsymbol{\theta})$ and non-smooth regularizer $\mathcal{R}(\boldsymbol{\theta})$
- the optimal solution is bounded, i.e. $\|\boldsymbol{\theta}^*\|_2 \leq D$
- each visited point is at a bounded distance from the optimal solution, i.e. $(\forall k)\ \|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^*\|_2 \leq D$
- both $\mathcal{L}$ and $\mathcal{R}$ are Lipschitz continuous, i.e. $\|\partial \mathcal{L}/\partial \boldsymbol{\theta}\|_2$, $\|\partial \mathcal{R}/\partial \boldsymbol{\theta}\|_2 \leq G$
- the non-smooth regularizer vanishes at zero, i.e. $\mathcal{R}(\mathbf{0}) = 0$

We additionally require that the errors do not change the Lipschitz continuity properties, i.e. $\|\partial \mathcal{L}/\partial \boldsymbol{\theta} + \boldsymbol{\xi}\|_2 \leq G$ (as discussed in Assumption 5.5).

5.3.1 Algorithm

We analyze *forward-backward splitting* [Duchi and Singer, 2009c] for deterministic as well as biased stochastic errors, for non-increasing step sizes of the form $\eta_k \in \mathcal{O}(\frac{1}{k^r})$ for $r > 0$. This method is equivalent to basic *proximal gradient* [Schmidt et al., 2011] for $r = 0$ (constant step size). We point out that FBS has $\mathcal{O}(\frac{1}{\sqrt{K}})$ convergence for $r = \frac{1}{2}$, while basic PG has $\mathcal{O}(\frac{1}{K})$ convergence, and accelerated PG has $\mathcal{O}(\frac{1}{K^2})$ convergence. Thus, PG methods have faster convergence but they are more sensitive to errors.

FBS performs gradient descent steps for the smooth part of the objective function, and (closed form) projection steps for the non-smooth part. Here we assume that at each iteration k, we approximate the gradient with some (deterministic or biased stochastic) error $\boldsymbol{\xi}^{(k)}$. For our objective function in eq.(5.2), one iteration of the algorithm is equivalent to:

$$\begin{aligned} &\text{i. } \boldsymbol{\theta}^{(k+\frac{1}{2})} = \boldsymbol{\theta}^{(k)} - \eta_k(\mathbf{g}_{\mathcal{L}}^{(k)} + \boldsymbol{\xi}^{(k)}) \\ &\text{ii. } \boldsymbol{\theta}^{(k+1)} \ = \arg\min_{\boldsymbol{\theta}}(\tfrac{1}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^{(k+\frac{1}{2})}\|_2^2 + \eta_{k+1}\mathcal{R}(\boldsymbol{\theta})) \end{aligned} \tag{5.8}$$

where $\mathbf{g}_{\mathcal{L}}^{(k)} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}^{(k)})$, and $\boldsymbol{\xi}^{(k)}$ is the error in the gradient approximation. Step ii is a projection step for the non-smooth regularizer $\mathcal{R}(\boldsymbol{\theta})$.

For the regularizer in our motivating problem $\mathcal{R}(\mathbf{W}) = \rho\|\mathbf{W}\|_1$, Step ii of eq.(5.8) decomposes into N^2 independent *lasso* problems [Tibshirani, 1996]. For clarity of exposition, we drop the subindices in the following equation. Let $w \equiv w_{n_1n_2}$, $w^{(k+\frac{1}{2})} \equiv w_{n_1n_2}^{(k+\frac{1}{2})}$, $w^{(k+1)} \equiv w_{n_1n_2}^{(k+1)}$, $\lambda \equiv \eta_{k+1}\rho$. The *lasso* problem for Step ii of eq.(5.8) is:

$$w^{(k+1)} = \arg\min_{w\in\mathbb{R}}\left(\frac{1}{2}(w - w^{(k+\frac{1}{2})})^2 + \lambda|w|\right) \tag{5.9}$$

for some $\lambda > 0$. The optimal solution of this problem is given by:

$$\begin{aligned} w^{(k+1)} &= \text{lasso}_\lambda(w^{(k+\frac{1}{2})}) \\ &= \text{sign}(w^{(k+\frac{1}{2})})\max(0, |w^{(k+\frac{1}{2})}| - \lambda) \end{aligned} \tag{5.10}$$

Note that eq.(5.10) can lead to sparse solutions, since whenever the absolute value of $w^{(k+\frac{1}{2})}$ is smaller than λ, the optimal solution $w^{(k+1)}$ is set to zero. Algorithm 5.1 shows the stochastic optimization method in detail.

Algorithm 5.1 Stochastic FOBOS for learning Ising models.

```
Input: empirical second-order moment $\widehat{\mathbf{\Sigma}}$, first-order moment $\widehat{\boldsymbol{\mu}}$, sparseness parameter $\rho > 0$
Initialize $\mathbf{W}^{(1)} = \mathbf{0}$, $\mathbf{b}^{(1)} = \mathbf{0}$
for each iteration $1, \dots, K$ do
    Use a $(B, V, S, D)$-sampler to produce biased estimates of the gradients $\partial \log \mathcal{Z}/\partial \mathbf{W} + \mathbf{\Xi}^{(k)}$ and $\partial \log \mathcal{Z}/\partial \mathbf{b} + \boldsymbol{\xi}^{(k)}$ with error terms $\mathbf{\Xi}^{(k)}$ and $\boldsymbol{\xi}^{(k)}$
    Update $\mathbf{W}^{(k+\frac{1}{2})} \leftarrow \mathbf{W}^{(k)} - \eta_k(\partial \log \mathcal{Z}/\partial \mathbf{W} + \mathbf{\Xi}^{(k)} - \widehat{\mathbf{\Sigma}})$
    Update $\mathbf{W}^{(k+1)} \leftarrow \mathbf{lasso}_{\eta_{k+1}\rho}(\mathbf{W}^{(k+\frac{1}{2})})$, where lasso is equivalent to apply eq.(5.10) entrywise
    Update $\mathbf{b}^{(k+1)} \leftarrow \mathbf{b}^{(k)} - \eta_k(\partial \log \mathcal{Z}/\partial \mathbf{b} + \boldsymbol{\xi}^{(k)} - \widehat{\boldsymbol{\mu}})$
end for
Output: the weighted average of all visited points $\frac{\sum_k \eta_k \mathbf{W}^{(k)}}{\sum_k \eta_k}$, $\frac{\sum_k \eta_k \mathbf{b}^{(k)}}{\sum_k \eta_k}$, the average of all visited points $\frac{\sum_k \mathbf{W}^{(k)}}{K}$, $\frac{\sum_k \mathbf{b}^{(k)}}{K}$, or the last visited point $\mathbf{W}^{(K)}$, $\mathbf{b}^{(K)}$
```

The following technical lemma is a building block for our analysis of convergence rates. The technical lemma generalizes Lemma 1 of [Duchi and Singer, 2009c], here we assume a sequence of deterministic errors.

Lemma 5.6. *For a sequence of deterministic errors $\boldsymbol{\xi}^{(1)}, \ldots, \boldsymbol{\xi}^{(K)}$ and non-increasing step sizes η_k, the objective function evaluated at each iteration is bounded as follows:*

$$\begin{aligned}&\eta_k(\mathcal{L}(\boldsymbol{\theta}^{(k)}) - \mathcal{L}(\boldsymbol{\theta}^*)) + \eta_{k+1}(\mathcal{R}(\boldsymbol{\theta}^{(k+1)}) - \mathcal{R}(\boldsymbol{\theta}^*))\\ &\le \tfrac{1}{2}\begin{pmatrix} \|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^*\|_2^2 - \|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^*\|_2^2 \\ +4D\eta_k\|\boldsymbol{\xi}^{(k)}\|_2 + 8\eta_k^2 G^2 \end{pmatrix}\end{aligned} \tag{5.11}$$

Proof. Let $\mathcal{L}^{(k)} \equiv \mathcal{L}(\boldsymbol{\theta}^{(k)})$, $\mathcal{R}^{(k)} \equiv \mathcal{R}(\boldsymbol{\theta}^{(k)})$, $\mathcal{L}^* \equiv \mathcal{L}(\boldsymbol{\theta}^*)$, $\mathcal{R}^* \equiv \mathcal{R}(\boldsymbol{\theta}^*)$ and $a^{(k)} \equiv \|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^*\|_2^2$.

As noted in [Duchi and Singer, 2009c], eq.(5.8) can be written as a single step:

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \eta_k(\mathbf{g}_{\mathcal{L}}^{(k)} + \boldsymbol{\xi}^{(k)}) - \eta_{k+1}\mathbf{g}_{\mathcal{R}}^{(k+1)} \tag{5.12}$$

where $\mathbf{g}_{\mathcal{R}}^{(k+1)} \in \frac{\partial \mathcal{R}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}^{(k+1)})$. This follows from the fact that $\boldsymbol{\theta}^{(k+1)}$ minimizes Step ii of eq.(5.8), if and only if $\mathbf{0}$ belongs to the subdifferential set of the non-smooth objective function evaluated at $\boldsymbol{\theta}^{(k+1)}$.

By eq.(5.12), $a^{(k+1)} = \|\boldsymbol{\theta}^{(k)} - \eta_k(\mathbf{g}_{\mathcal{L}}^{(k)} + \boldsymbol{\xi}^{(k)}) - \eta_{k+1}\mathbf{g}_{\mathcal{R}}^{(k+1)} - \boldsymbol{\theta}^*\|_2^2 = a^{(k)} + 2\eta_k F_1 + 2\eta_{k+1}F_2 + 2\eta_{k+1}F_3 + 2\eta_k F_4 + F_5$ for $F_1 \equiv -\langle \mathbf{g}_{\mathcal{L}}^{(k)}, \boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^*\rangle$, $F_2 \equiv -\langle \mathbf{g}_{\mathcal{R}}^{(k+1)}, \boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^*\rangle$, $F_3 \equiv \langle \mathbf{g}_{\mathcal{R}}^{(k+1)}, \boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}\rangle$, $F_4 \equiv -\langle \boldsymbol{\xi}^{(k)}, \boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^*\rangle$ and $F_5 \equiv \|\eta_k(\mathbf{g}_{\mathcal{L}}^{(k)} + \boldsymbol{\xi}^{(k)}) + \eta_{k+1}\mathbf{g}_{\mathcal{R}}^{(k+1)}\|_2^2$.

By the definition of subgradients of convex functions, $F_1 \le \mathcal{L}^* - \mathcal{L}^{(k)}$ and $F_2 \le \mathcal{R}^* - \mathcal{R}^{(k+1)}$.

By eq.(5.12), the Cauchy-Schwarz inequality and Assumption 5.5, $F_3 = \langle \mathbf{g}_{\mathcal{R}}^{(k+1)}, -\eta_k(\mathbf{g}_{\mathcal{L}}^{(k)} + \boldsymbol{\xi}^{(k)}) - \eta_{k+1}\mathbf{g}_{\mathcal{R}}^{(k+1)}\rangle \le \|\mathbf{g}_{\mathcal{R}}^{(k+1)}\|_2 \|\eta_k(\mathbf{g}_{\mathcal{L}}^{(k)} + \boldsymbol{\xi}^{(k)}) + \eta_{k+1}\mathbf{g}_{\mathcal{R}}^{(k+1)}\|_2 \le \|\mathbf{g}_{\mathcal{R}}^{(k+1)}\|_2(\eta_k\|\mathbf{g}_{\mathcal{L}}^{(k)} + \boldsymbol{\xi}^{(k)}\|_2 + \eta_{k+1}\|\mathbf{g}_{\mathcal{R}}^{(k+1)}\|_2) \le (\eta_k + \eta_{k+1})G^2$.

By the Cauchy-Schwarz inequality, $F_4 \le \|\boldsymbol{\xi}^{(k)}\|_2 \|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^*\|_2 \le D\|\boldsymbol{\xi}^{(k)}\|_2$, since by assumption $(\forall k)\ a^{(k)} \le D^2$.

By the Cauchy-Schwarz inequality and Assumption 5.5, $F_5 \le \eta_k^2\|(\mathbf{g}_{\mathcal{L}}^{(k)} + \boldsymbol{\xi}^{(k)})\|_2^2 + 2\eta_k\eta_{k+1}\langle \mathbf{g}_{\mathcal{L}}^{(k)} + \boldsymbol{\xi}^{(k)}, \mathbf{g}_{\mathcal{R}}^{(k+1)}\rangle + \eta_{k+1}^2\|\mathbf{g}_{\mathcal{R}}^{(k+1)}\|_2^2 \le \eta_k^2\|(\mathbf{g}_{\mathcal{L}}^{(k)} + \boldsymbol{\xi}^{(k)})\|_2^2 + 2\eta_k\eta_{k+1}\|\mathbf{g}_{\mathcal{L}}^{(k)} + \boldsymbol{\xi}^{(k)}\|_2\|\mathbf{g}_{\mathcal{R}}^{(k+1)}\|_2 + \eta_{k+1}^2\|\mathbf{g}_{\mathcal{R}}^{(k+1)}\|_2^2 \le (\eta_k^2 + 2\eta_k\eta_{k+1} + \eta_{k+1}^2)G^2$.

Putting everything together, $a^{(k+1)} \le a^{(k)} + 2\eta_k(\mathcal{L}^* - \mathcal{L}^{(k)}) + 2\eta_{k+1}(\mathcal{R}^* - \mathcal{R}^{(k+1)}) + 2\eta_k D\|\boldsymbol{\xi}^{(k)}\|_2 + (\eta_k^2 + 4\eta_k\eta_{k+1} + 3\eta_{k+1}^2)G^2$. Finally, since $\eta_{k+1} \le \eta_k \Rightarrow (\eta_k^2 + 4\eta_k\eta_{k+1} + 3\eta_{k+1}^2)G^2 \le 8\eta_k^2 G^2$. □

5.3.2 Convergence Rates for Deterministic Errors

In what follows, we analyze three different flavors of forward-backward splitting: *robust* which outputs the weighted average of all visited points by using the step sizes as in *robust stochastic approximation* [Nemirovski et al., 2009], *basic* which outputs the average of all visited points as in [Duchi and Singer, 2009c], or *random* which outputs a point chosen uniformly at random from the visited points. Here we assume that at each iteration k, we approximate the gradient with some deterministic error $\boldsymbol{\xi}^{(k)}$. Our results in this subsection will allow us to draw some conclusions regarding not only FBS but also proximal gradient.

Note that in our case, proving convergence of the best visited point as in the *subgradient method* [Shor, 1985] or in the *proximal-gradient method* [Schmidt et al., 2011] (i.e. $\boldsymbol{\theta}^{(k^*)}$, $k^* = \arg\min_k(\mathcal{L}(\boldsymbol{\theta}^{(k)}) + \mathcal{R}(\boldsymbol{\theta}^{(k)}))$) is not useful, since computing the partition function is NP-hard. Despite this fact, such proof is elementary since the minimum value of the objective function from all visited points is less than or equal to the average (or weighted average) of all visited points. Therefore, the bounds in Theorems 5.7 and 5.8 are also bounds for the convergence of the best visited point.

In order to make our bounds more general for different choices of step size $\eta_k \in \mathcal{O}(\frac{1}{k^r})$ for some $r > 0$, we use *generalized harmonic numbers* $H_{r,K} = \sum_{k=1}^{K} \frac{1}{k^r}$ and therefore $H_{0,K} = K$, $H_{1,K} \approx \log K$, $H_{r,K} \approx \frac{K^{1-r}}{1-r}$ for $0 < r < 1$ and $H_{r,K} \approx \frac{1-K^{1-r}}{r-1}$ for $r > 1$.

Additionally, we define a weighted error term that will be used for our analysis of deterministic as well as biased stochastic errors. Given a sequence of errors $\boldsymbol{\xi}^{(1)}, \dots, \boldsymbol{\xi}^{(K)}$ and a set of arbitrary weights γ_k such that $\sum_k \gamma_k = 1$, the error term is defined as:

$$A_{\boldsymbol{\gamma},\boldsymbol{\xi}} \equiv \textstyle\sum_k \gamma_k \|\boldsymbol{\xi}^{(k)}\|_2 \tag{5.13}$$

First, we show the convergence rate of robust FBS.

Theorem 5.7. *For a sequence of deterministic errors $\boldsymbol{\xi}^{(1)}, \dots, \boldsymbol{\xi}^{(K)}$, step size $\eta_k = \frac{\beta}{Gk^r}$ for $r > 0$, initial point $\boldsymbol{\theta}^{(1)} = \mathbf{0}$, the objective function evaluated at the weighted average of all visited points converges to the optimal solution with rate:*

$$\begin{aligned} &\mathcal{L}(\overline{\boldsymbol{\theta}}) + \mathcal{R}(\overline{\boldsymbol{\theta}}) - \mathcal{L}(\boldsymbol{\theta}^*) - \mathcal{R}(\boldsymbol{\theta}^*) \leq \pi_\eta(K) \\ &\leq \tfrac{D^2 G}{2\beta H_{r,K}} + 2DA_{\boldsymbol{\gamma},\boldsymbol{\xi}} + \tfrac{4\beta G H_{2r,K}}{H_{r,K}} \end{aligned} \tag{5.14}$$

where $\overline{\boldsymbol{\theta}} = \frac{\sum_k \eta_k \boldsymbol{\theta}^{(k)}}{\sum_k \eta_k}$, the weighted average regret $\pi_\eta(K) = \frac{\sum_k \eta_k (\mathcal{L}(\boldsymbol{\theta}^{(k)}) + \mathcal{R}(\boldsymbol{\theta}^{(k)}))}{\sum_k \eta_k} - \mathcal{L}(\boldsymbol{\theta}^*) - \mathcal{R}(\boldsymbol{\theta}^*)$, the error term $A_{\boldsymbol{\gamma},\boldsymbol{\xi}}$ is defined as in eq.(5.13), and the error weights $\gamma_k = \frac{1/k^r}{H_{r,K}}$ such that $\sum_k \gamma_k = 1$.

Proof. Let $\mathcal{L}^{(k)} \equiv \mathcal{L}(\boldsymbol{\theta}^{(k)})$, $\mathcal{R}^{(k)} \equiv \mathcal{R}(\boldsymbol{\theta}^{(k)})$, $\mathcal{L}^* \equiv \mathcal{L}(\boldsymbol{\theta}^*)$, $\mathcal{R}^* \equiv \mathcal{R}(\boldsymbol{\theta}^*)$ and $a^{(k)} \equiv \|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^*\|_2^2$.

By Jensen's inequality $\mathcal{L}(\overline{\boldsymbol{\theta}}) + \mathcal{R}(\overline{\boldsymbol{\theta}}) \leq \sum_k \eta_k (\mathcal{L}^{(k)} + \mathcal{R}^{(k)}) / \sum_k \eta_k$. Therefore $\mathcal{L}(\overline{\boldsymbol{\theta}}) - \mathcal{L}^* + \mathcal{R}(\overline{\boldsymbol{\theta}}) - \mathcal{R}^* \leq \pi_\eta(K) \leq (\eta_1 \mathcal{R}^{(1)} + \sum_k (\eta_k (\mathcal{L}^{(k)} - \mathcal{L}^*) + \eta_{k+1}(\mathcal{R}^{(k+1)} - \mathcal{R}^*))) / \sum_k \eta_k \equiv F$, and since $\boldsymbol{\theta}^{(1)} = \mathbf{0} \Rightarrow \mathcal{R}^{(1)} = 0$.

By Lemma 5.6 we know that $\eta_k (\mathcal{L}^{(k)} - \mathcal{L}^*) + \eta_{k+1}(\mathcal{R}^{(k+1)} - \mathcal{R}^*) \leq \frac{1}{2}(a^{(k)} - a^{(k+1)} + 4D\eta_k \|\boldsymbol{\xi}^{(k)}\|_2 + 8\eta_k^2 G^2) \Rightarrow (\sum_k \eta_k) F \leq \frac{1}{2} \sum_k (a^{(k)} - a^{(k+1)}) + 2D(\sum_k \eta_k \|\boldsymbol{\xi}^{(k)}\|_2) + 4(\sum_k \eta_k^2) G^2 \leq \frac{a^{(1)}}{2} + 2D(\sum_k \eta_k \|\boldsymbol{\xi}^{(k)}\|_2) + 4(\sum_k \eta_k^2) G^2$.

Since by assumption $(\forall k)\ a^{(k)} \leq D^2 \Rightarrow (\sum_k \eta_k) F \leq \frac{D^2}{2} + 2D(\sum_k \eta_k \|\boldsymbol{\xi}^{(k)}\|_2) + 4(\sum_k \eta_k^2) G^2$. Finally, by replacing $\eta_k = \frac{\beta}{Gk^r}$, we prove our claim. □

Second, we show the convergence rate of basic FBS.

Theorem 5.8. *For a sequence of deterministic errors $\boldsymbol{\xi}^{(1)}, \dots, \boldsymbol{\xi}^{(K)}$, step size $\eta_k = \frac{\beta}{Gk^r}$ for $r > 0$, initial point $\boldsymbol{\theta}^{(1)} = \mathbf{0}$, the objective function evaluated at the average of all visited points converges to the optimal solution with rate:*

$$\begin{aligned} &\mathcal{L}(\overline{\boldsymbol{\theta}}) + \mathcal{R}(\overline{\boldsymbol{\theta}}) - \mathcal{L}(\boldsymbol{\theta}^*) - \mathcal{R}(\boldsymbol{\theta}^*) \leq \pi(K) \\ &\leq \tfrac{D^2 G (K+1)^r}{2\beta K} + 2^{1+r} DA_{\boldsymbol{\gamma},\boldsymbol{\xi}} + \tfrac{2^{2+r}\beta G H_{r,K}}{K} \end{aligned} \tag{5.15}$$

where $\overline{\boldsymbol{\theta}} = \frac{\sum_k \boldsymbol{\theta}^{(k)}}{K}$, the average regret $\pi(K) = \frac{\sum_k (\mathcal{L}(\boldsymbol{\theta}^{(k)}) + \mathcal{R}(\boldsymbol{\theta}^{(k)}))}{K} - \mathcal{L}(\boldsymbol{\theta}^*) - \mathcal{R}(\boldsymbol{\theta}^*)$, the error term $A_{\boldsymbol{\gamma},\boldsymbol{\xi}}$ is defined as in eq.(5.13), and the error weights $\gamma_k = \frac{1}{K}$ such that $\sum_k \gamma_k = 1$.

Proof. Let $\mathcal{L}^{(k)} \equiv \mathcal{L}(\boldsymbol{\theta}^{(k)})$, $\mathcal{R}^{(k)} \equiv \mathcal{R}(\boldsymbol{\theta}^{(k)})$, $\mathcal{L}^* \equiv \mathcal{L}(\boldsymbol{\theta}^*)$, $\mathcal{R}^* \equiv \mathcal{R}(\boldsymbol{\theta}^*)$ and $a^{(k)} \equiv \|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^*\|_2^2$.

By Jensen's inequality $\mathcal{L}(\overline{\boldsymbol{\theta}}) + \mathcal{R}(\overline{\boldsymbol{\theta}}) \leq \sum_k (\mathcal{L}^{(k)} + \mathcal{R}^{(k)})/K$. Therefore $\mathcal{L}(\overline{\boldsymbol{\theta}}) - \mathcal{L}^* + \mathcal{R}(\overline{\boldsymbol{\theta}}) - \mathcal{R}^* \leq \pi(K) \leq (\mathcal{R}^{(1)} + \sum_k (\mathcal{L}^{(k)} - \mathcal{L}^* + \mathcal{R}^{(k+1)} - \mathcal{R}^*))/K \equiv F$, and since $\boldsymbol{\theta}^{(1)} = \mathbf{0} \Rightarrow \mathcal{R}^{(1)} = 0$.

For using Lemma 5.6, note that since $\eta_{k+1} \leq \eta_k \Rightarrow \eta_{k+1}(\mathcal{L}^{(k)} - \mathcal{L}^* + \mathcal{R}^{(k+1)} - \mathcal{R}^*) \leq \eta_k(\mathcal{L}^{(k)} - \mathcal{L}^*) + \eta_{k+1}(\mathcal{R}^{(k+1)} - \mathcal{R}^*) \leq \frac{1}{2}(a^{(k)} - a^{(k+1)} + 4D\eta_k\|\boldsymbol{\xi}^{(k)}\|_2 + 8\eta_k^2 G^2)$. Furthermore, since $\frac{\eta_k}{\eta_{k+1}} \leq 2^r \Rightarrow KF \leq \frac{1}{2}\sum_k \frac{a^{(k)} - a^{(k+1)}}{\eta_{k+1}} + 2^{1+r}D(\sum_k \|\boldsymbol{\xi}^{(k)}\|_2) + 2^{2+r}(\sum_k \eta_k)G^2 \leq \frac{a^{(1)}}{2\eta_2} + \frac{1}{2}\sum_{k=2}^{K}\left(\frac{a^{(k)}}{\eta_{k+1}} - \frac{a^{(k)}}{\eta_k}\right) + 2^{1+r}D(\sum_k \|\boldsymbol{\xi}^{(k)}\|_2) + 2^{2+r}(\sum_k \eta_k)G^2$.

Since by assumption $(\forall k)\ a^{(k)} \leq D^2$ we have $KF \leq \frac{D^2}{2}\left(\frac{1}{\eta_2} + \sum_{k=2}^{K}\left(\frac{1}{\eta_{k+1}} - \frac{1}{\eta_k}\right)\right) + 2^{1+r}D(\sum_k \|\boldsymbol{\xi}^{(k)}\|_2) + 2^{2+r}(\sum_k \eta_k)G^2 \leq \frac{D^2}{2\eta_{K+1}} + 2^{1+r}D(\sum_k \|\boldsymbol{\xi}^{(k)}\|_2) + 2^{2+r}(\sum_k \eta_k)G^2$. Finally, by replacing $\eta_k = \frac{\beta}{Gk^r}$, we prove our claim. □

Finally, we show the convergence rate of random FBS.

Theorem 5.9. *For a sequence of deterministic errors $\boldsymbol{\xi}^{(1)}, \dots, \boldsymbol{\xi}^{(K)}$, step size $\eta_k = \frac{\beta}{Gk^r}$ for $r > 0$, initial point $\boldsymbol{\theta}^{(1)} = \mathbf{0}$ and some confidence parameter $0 < \varepsilon < 1$, the objective function evaluated at a point k chosen uniformly at random from the visited points converges, with probability at least $1 - \varepsilon$, to the optimal solution with rate:*

$$\begin{aligned} &\mathcal{L}(\boldsymbol{\theta}^{(k)}) + \mathcal{R}(\boldsymbol{\theta}^{(k)}) - \mathcal{L}(\boldsymbol{\theta}^*) - \mathcal{R}(\boldsymbol{\theta}^*) \\ &\leq \tfrac{1}{\varepsilon}\left(\tfrac{D^2 G(K+1)^r}{2\beta K} + 2^{1+r}DA_{\boldsymbol{\gamma},\boldsymbol{\xi}} + \tfrac{2^{2+r}\beta G H_{r,K}}{K}\right) \end{aligned} \tag{5.16}$$

where the error term $A_{\boldsymbol{\gamma},\boldsymbol{\xi}}$ is defined as in eq.(5.13)*, and the error weights $\gamma_k = \frac{1}{K}$ such that $\sum_k \gamma_k = 1$.*

Proof. Let $\mathcal{L}^{(k)} \equiv \mathcal{L}(\boldsymbol{\theta}^{(k)})$, $\mathcal{R}^{(k)} \equiv \mathcal{R}(\boldsymbol{\theta}^{(k)})$, $\mathcal{L}^* \equiv \mathcal{L}(\boldsymbol{\theta}^*)$, $\mathcal{R}^* \equiv \mathcal{R}(\boldsymbol{\theta}^*)$ and $\mathcal{U}$ the uniform distribution for $k \in \{1, \dots, K\}$.

By Markov's inequality, for $a^{(k)} = \mathcal{L}^{(k)} + \mathcal{R}^{(k)} - \mathcal{L}^* - \mathcal{R}^* \geq 0$, we have $\mathbb{P}_{\mathcal{U}}[a^{(k)} \geq c] \leq \frac{\mathbb{E}_{\mathcal{U}}[a^{(k)}]}{c}$. Note that $\mathbb{E}_{\mathcal{U}}[a^{(k)}] = \frac{1}{K}\sum_k (\mathcal{L}^{(k)} + \mathcal{R}^{(k)}) - \mathcal{L}^* - \mathcal{R}^* = \pi(K)$. By Theorem 5.8, we know that $\pi(K) \leq \frac{D^2 G(K+1)^r}{2\beta K} + 2^{1+r}DA_{\boldsymbol{\gamma},\boldsymbol{\xi}} + \frac{2^{2+r}\beta G H_{r,K}}{K} \equiv F$, therefore $\mathbb{P}_{\mathcal{U}}[a^{(k)} \geq c] \leq \frac{F}{c}$. For $c = \frac{F}{\varepsilon} \Rightarrow \mathbb{P}_{\mathcal{U}}[a^{(k)} \geq \frac{F}{\varepsilon}] \leq \varepsilon$. □

The convergence rates in Theorems 5.7, 5.8 and 5.9 lead to an error term $A_{\boldsymbol{\gamma},\boldsymbol{\xi}}$ that is linear, while the error term is quadratic in the analysis of proximal gradient [Schmidt et al., 2011]. In basic PG, the error term can be written as:

$$\tfrac{1}{K}(\textstyle\sum_k \|\boldsymbol{\xi}^{(k)}\|_2)^2 = K(A_{\boldsymbol{\gamma},\boldsymbol{\xi}})^2 \tag{5.17}$$

where the error weights $\gamma_k = \frac{1}{K}$ such that $\sum_k \gamma_k = 1$. In accelerated PG, the error term can be written as:

$$\tfrac{4}{(K+1)^2}(\textstyle\sum_k k\|\boldsymbol{\xi}^{(k)}\|_2)^2 = K^2(A_{\boldsymbol{\gamma},\boldsymbol{\xi}})^2 \tag{5.18}$$

where the error weights $\gamma_k = k/\binom{K}{2}$ so that $\sum_k \gamma_k = 1$.

Note that both PG methods contain terms K and K^2, which are not in our analysis. As noted in [Schmidt et al., 2011], errors have a greater effect on the accelerated method

Table 5.1: Order of errors $\|\boldsymbol{\xi}^{(k)}\|_2$ required to obtain convergence of the error term for the *deterministic* case: basic (PB) and accelerated (PA) proximal gradient, basic (FB) and robust (FR) forward-backward splitting.

Method	**Convergence**			
	for $K\to+\infty$	$\mathcal{O}(\frac{1}{\sqrt{K}})$	$\mathcal{O}(\frac{1}{K})$	$\mathcal{O}(\frac{1}{K^2})$
PB	$\mathcal{O}(\frac{1}{k^{1/2+\epsilon}})$	$\mathcal{O}(\frac{1}{k^{3/4+\epsilon}})$	$\mathcal{O}(\frac{1}{k^{1+\epsilon}})$	-
PA	$\mathcal{O}(\frac{1}{k^{1+\epsilon}})$	$\mathcal{O}(\frac{1}{k^{5/4+\epsilon}})$	$\mathcal{O}(\frac{1}{k^{3/2+\epsilon}})$	$\mathcal{O}(\frac{1}{k^{2+\epsilon}})$
FB $(r=\frac{1}{2})$	$\mathcal{O}(\frac{1}{\log k})$	$\mathcal{O}(\frac{1}{k^{1/2+\epsilon}})$	$\mathcal{O}(\frac{1}{k^{1+\epsilon}})$	-
FR $(r=\frac{1}{2})$	$\mathcal{O}(\frac{1}{\log k})$	$\mathcal{O}(\frac{1}{k^{1/2+\epsilon}})$	-	-

than on the basic method. This observation suggests that, unlike in the error-free case, accelerated PG is not necessarily better than the basic method due to a higher sensitivity to errors [Devolder et al., 2011].

Intuitively speaking, basic PG is similar to basic FBS in the sense that errors from all iterations have the same effect on the convergence rate, i.e. γ_k is constant. In robust FBS, errors in the last iterations have a lower effect on the convergence rate than errors in the beginning, i.e. γ_k is decreasing. In accelerated PG, errors in the last iterations have a bigger effect on the convergence rate than errors in the beginning, i.e. γ_k is increasing.

The analysis of Schmidt et al. [2011] for deterministic errors implies that in order to have convergence, the errors must decrease at a rate $\|\boldsymbol{\xi}^{(k)}\|_2 \in \mathcal{O}(\frac{1}{k^{1/2+\epsilon}})$ for some $\epsilon > 0$ in the case of basic PG, and $\mathcal{O}(\frac{1}{k^{1+\epsilon}})$ for accelerated PG. In contrast, our analysis of FBS show that we only need logarithmically decreasing errors $\mathcal{O}(\frac{1}{\log k})$ in order to have convergence. Regarding $\mathcal{O}(\frac{1}{\sqrt{K}})$ convergence of the error term $A_{\boldsymbol{\gamma},\boldsymbol{\xi}}$, basic and robust FBS requires errors $\mathcal{O}(\frac{1}{k^{1/2+\epsilon}})$ (the minimum required for convergence in basic PG). Table 5.1 summarizes the requirements for different convergence rates of the error term $A_{\boldsymbol{\gamma},\boldsymbol{\xi}}$ of FBS as well as the error terms of basic PG in eq.(5.17) and accelerated PG in eq.(5.18).

For an informal (and incomplete) analysis of the results in [Schmidt et al., 2011] for biased stochastic optimization, consider each error bounded by its bias and variance $\|\boldsymbol{\xi}^{(k)}\|_2 \leq B/S_k + c\sqrt{V/S_k}$ for some $c > 0$ and an increasing number of random samples S_k that allows to obtain decreasing errors. Without noting the possible need of "uniform convergence" of the bound for all K iterations (making c a function of K), the number of random samples must increase (at least) at a rate that is quadratic of the rate of the errors. For instance, in order to have $\mathcal{O}(\frac{1}{K})$ convergence, basic PG requires errors to be $\mathcal{O}(\frac{1}{k^{1+\epsilon}})$ and therefore it would require (at least) an increasing number of random samples $S_k \in \mathcal{O}(k^{2+\epsilon})$ for some $\epsilon > 0$. Accelerated PG would require (at least) $S_k \in \mathcal{O}(k^{4+\epsilon})$ in order to obtain $\mathcal{O}(\frac{1}{K^2})$ convergence. If we include the fact that c is a function of K, then the required number of random samples would be "worse than quadratic" of the required rate of the errors. Fortunately, a formal analysis in the next subsection shows that this is not the case for all methods except accelerated PG.

5.3.3 Bounding the Error Term for Biased Stochastic Optimization

In what follows, we focus in the analysis of stochastic errors in order to see if better convergence rates can be obtained than the ones informally outlined in the previous subsection. A formal analysis of the error terms show that *forward-backward splitting* for biased stochastic errors requires only a logarithmically increasing number of random samples in order to converge, i.e. $S_k \in \mathcal{O}(\log k)$. More interestingly, we found that the required number of random samples is the same for the deterministic and the biased stochastic setting for FBS and basic PG. On the negative side, we found that accelerated PG is not guaranteed to converge in the biased stochastic setting.

Next, we present our high probability bound for the error term for biased stochastic optimization. One way to bound the error term $A_{\boldsymbol{\gamma},\boldsymbol{\xi}}$ would be to rely on "uniform convergence" arguments, i.e. to bound the error of each iteration $\|\boldsymbol{\xi}^{(k)}\|_2$ and then use the well-known union bound. We chose to bound the error term itself, by using the fact that errors become independent (but not identically distributed) conditioned to the parameters $\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(K)}$. We also allow for a different number of random samples S_k for each iteration k.

Theorem 5.10. *Given K (B, V, S_k, D)-samplers each producing estimates with an error $\boldsymbol{\xi}^{(k)}$, and given a set of arbitrary weights γ_k such that $\sum_k \gamma_k = 1$. For some confidence parameter $0 < \delta < 1$, with probability at least $1 - \delta$, the error term is bounded as follows:*

$$A_{\boldsymbol{\gamma},\boldsymbol{\xi}} \leq \lambda_1 + \tfrac{2\sqrt{M}}{3K} \log \tfrac{1}{\delta} + \sqrt{2\lambda_2 \log \tfrac{1}{\delta} + \tfrac{4M}{9K^2} \log^2 \tfrac{1}{\delta}} \tag{5.19}$$

where the bias term $\lambda_1 = \min(2\sqrt{M}, B \sum_k \frac{\gamma_k}{S_k})$ and the variance term $\lambda_2 = \min(4M, V \sum_k \frac{\gamma_k^2}{S_k})$.

Proof. Let $\mathcal{Q}_k$ be the distribution of the error for the k-th sampler, the joint distribution $\mathcal{Q} \equiv \{\mathcal{Q}_1, \dots, \mathcal{Q}_K\}$, $\mathcal{T}$ be the joint distribution of $\boldsymbol{\Theta} \equiv \{\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(K)}\}$, the first-order moment $\phi_k \equiv \mathbb{E}_{\mathcal{Q}}[\|\boldsymbol{\xi}^{(k)}\|_2]$ and the second-order moment $\nu_k^2 \equiv \mathbb{V}\mathrm{ar}_{\mathcal{Q}}[\|\boldsymbol{\xi}^{(k)}\|_2]$.

By Lemma 5.3 we know that $\|\partial \log \mathcal{Z}/\partial\boldsymbol{\theta}\|_\infty \leq 1$. By Assumption 5.5, for any sampler we have $\|\partial \log \mathcal{Z}/\partial\boldsymbol{\theta} + \boldsymbol{\xi}^{(k)}\|_\infty \leq 1$ and therefore $\|\boldsymbol{\xi}^{(k)}\|_\infty \leq 2$ in the worst case. Therefore $\|\boldsymbol{\xi}^{(k)}\|_2 \leq \sqrt{M}\|\boldsymbol{\xi}^{(k)}\|_\infty \leq 2\sqrt{M}$.

Given that the error is bounded, we have $\mathbb{E}_{\mathcal{Q}}[\|\boldsymbol{\xi}^{(k)}\|_2] \leq 2\sqrt{M}$. By using the bounds in Definition 5.4, the bias is at most $\phi_k \leq \min(2\sqrt{M}, \frac{B}{S_k})$.

Similarly, we have $\mathbb{V}\mathrm{ar}_{\mathcal{Q}}[\|\boldsymbol{\xi}^{(k)}\|_2] = \mathbb{E}_{\mathcal{Q}}[\|\boldsymbol{\xi}^{(k)}\|_2^2] - \mathbb{E}_{\mathcal{Q}}[\|\boldsymbol{\xi}^{(k)}\|_2]^2 \leq \mathbb{E}_{\mathcal{Q}}[\|\boldsymbol{\xi}^{(k)}\|_2^2] \leq 4M$. By using the bounds in Definition 5.4, the variance is at most $\nu_k^2 \leq \min(4M, \frac{V}{S_k})$.

Consider the variable $z_k = K\gamma_k\|\boldsymbol{\xi}^{(k)}\|_2$. Note that the mean $\widehat{z} = \frac{1}{K}\sum_k z_k = \sum_k \gamma_k \|\boldsymbol{\xi}^{(k)}\|_2 = A_{\boldsymbol{\gamma},\boldsymbol{\xi}}$ is the expression we want to upper-bound. The expected value $\overline{\phi} = \mathbb{E}_{\mathcal{Q}}[\widehat{z}] = \sum_k \gamma_k \phi_k \leq \min(2\sqrt{M}, B\sum_k \frac{\gamma_k}{S_k}) \equiv \lambda_1$. The average variance $\sigma^2 = \frac{1}{K}\sum_k \mathbb{V}\mathrm{ar}_{\mathcal{Q}}[z_k] = K\sum_k \gamma_k^2 \nu_k^2 \leq K\min(4M\sum_k \gamma_k^2, V\sum_k \frac{\gamma_k^2}{S_k}) \leq K\min(4M, V\sum_k \frac{\gamma_k^2}{S_k}) \equiv K\lambda_2$.

Our goal is to find an upper bound for $F_1 \equiv \mathbb{P}_{\mathcal{Q}}[\widehat{z} \geq \lambda_1 + \epsilon]$. By the definition of marginal distribution $F_1 = \int_{\boldsymbol{\Theta}} \mathbb{P}_{\mathcal{Q}}[\widehat{z} \geq \lambda_1 + \epsilon \mid \boldsymbol{\Theta}] p_{\mathcal{T}}(\boldsymbol{\Theta}) \leq \int_{\boldsymbol{\Theta}} \mathbb{P}_{\mathcal{Q}}[\widehat{z} \geq \overline{\phi} + \epsilon \mid \boldsymbol{\Theta}] p_{\mathcal{T}}(\boldsymbol{\Theta}) \equiv F_2$. By using the Bernstein inequality, $F_2 \leq \int_{\boldsymbol{\Theta}} e^{-\frac{K\epsilon^2}{2\sigma^2 + 4\sqrt{M}\epsilon/3}} p_{\mathcal{T}}(\boldsymbol{\Theta}) \leq \int_{\boldsymbol{\Theta}} e^{-\frac{K\epsilon^2}{2K\lambda_2 + 4\sqrt{M}\epsilon/3}} p_{\mathcal{T}}(\boldsymbol{\Theta}) =$

Table 5.2: Random samples S_k required to obtain convergence of the error term for the *biased stochastic* case: basic (PB) and accelerated (PA) proximal gradient, basic (FB) and robust (FR) forward-backward splitting.

Method	**Convergence**			
	for $K\to+\infty$	$\mathcal{O}(\frac{1}{\sqrt{K}})$	$\mathcal{O}(\frac{1}{K})$	$\mathcal{O}(\frac{1}{K^2})$
PB	$\mathcal{O}(k^{1/2+\epsilon})$	$\mathcal{O}(k^{3/4+\epsilon})$	$\mathcal{O}(k^{1+\epsilon})$	-
PA	-	-	-	-
FB $(r=\frac{1}{2})$	$\mathcal{O}(\log k)$	$\mathcal{O}(k^{1/2+\epsilon})$	$\mathcal{O}(k^{1+\epsilon})$	-
FR $(r=\frac{1}{2})$	$\mathcal{O}(\log k)$	$\mathcal{O}(k^{1/2+\epsilon})$	-	-

$e^{-\frac{K\epsilon^2}{2K\lambda_2+4\sqrt{M}\epsilon/3}}\int_{\mathbf{\Theta}} p_{\mathcal{T}}(\mathbf{\Theta}) = e^{-\frac{K\epsilon^2}{2K\lambda_2+4\sqrt{M}\epsilon/3}} = \delta$. By solving for ϵ in the last equality, we prove our claim. □

It is interesting to note what happens for a fixed number of random samples $S_k \in \mathcal{O}(1)$. In this case, the bias term $\lambda_1 \in \mathcal{O}(1)$ and therefore FBS will not converge. For robust FBS, the variance term $\lambda_2 \in \mathcal{O}(H_{2r,K}/(H_{r,K})^2)$ which for instance for $r = \frac{1}{2}$ we have $\lambda_2 \in \mathcal{O}(\frac{\log K}{K})$. For basic FBS, the variance term $\lambda_2 \in \mathcal{O}(\frac{1}{K})$. Therefore, for the constant number of random samples, the lack of convergence of FBS is explained only by the bias of the sampler and not its variance.

Table 5.2 summarizes the requirements for different convergence rates of the error term $A_{\boldsymbol{\gamma},\boldsymbol{\xi}}$ of FBS as well as the error terms of basic PG in eq.(5.17) and accelerated PG in eq.(5.18). Note that convergence for FBS is guaranteed for a logarithmically increasing number of random samples $S_k \in \mathcal{O}(\log k)$. Moreover, in order to obtain convergence rates of $\mathcal{O}(\frac{1}{\sqrt{K}})$ and $\mathcal{O}(\frac{1}{K})$, the required number of random samples is just the inverse of the required rate of the errors for the deterministic case, and not "worse than quadratic" as outlined in our informal analysis of the previous subsection.

One important conclusion from Theorem 5.10 is that the upper bound of the error term is $\Omega(\frac{1}{K})$ independently of the bias term λ_1 and the variance term λ_2. This implies that the error term is $\mathcal{O}(\frac{1}{K})$ for any setting of error weights γ_k and number of random samples S_k. The main implication is that the error term in accelerated PG in eq.(5.18) is constant and therefore the accelerated method is not guaranteed to converge.

5.4 Experimental Results

We illustrate our theoretical findings with a small synthetic experiment ($N = 15$ variables) since we want to report the log-likelihood at each iteration. We performed 10 repetitions. For each repetition, we generate edges in the ground truth model $\mathbf{W}_g$ with a 50% density. The weight of each edge is generated uniformly at random from $[-1; +1]$. We set $\mathbf{b}_g = \mathbf{0}$. We finally generate a dataset of 50 samples. We used a "Gibbs sampler" by first finding the mean field distribution and then performing 5 Gibbs iterations. We used a step size factor $\beta = 1$ and regularization parameter $\rho = 1/16$. We also include a two-step algorithm, by first learning the structure by ℓ_1-regularized logistic regression [Wainwright et al., 2006] and then

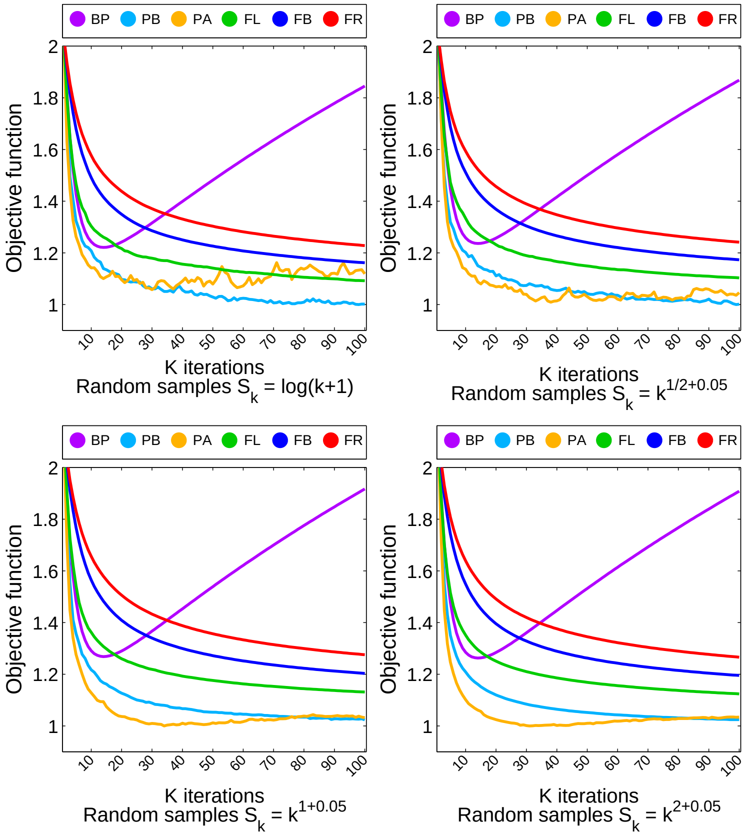


Figure 5.1: Objective function for different settings of increasing number of random samples. Basic (PB) and accelerated (PA) are noisier and require more samples than last point (FL), basic (FB) and robust (FR) forward-backward splitting in order to converge, but they exhibit faster convergence. Belief propagation (BP) does not converge.

learning the parameters by using FBS with belief propagation for gradient approximation. We summarize our results in Figure 5.1.

Our experiments suggest that stochastic optimization converges to the maximum likelihood estimate. We also show the Kullback-Leibler divergence to the ground truth, and more pronounced effects for importance sampling (Please, see Appendix F).

5.5 Concluding Remarks

There are several ways of extending this research. Although we focused on Ising models, the ideas developed in the current chapter could be applied to Markov random fields with higher order cliques. Our analysis can be easily extended to parameter learning for fixed structures by using a ℓ_2^2 regularizer instead. Although we show that accelerated proximal gradient is not guaranteed to converge in our specific biased stochastic setting, necessary conditions for its convergence needs to be investigated.

Chapter 6

Lipschitz Parameterization of Probabilistic Graphical Models

In the previous chapters, we proposed priors and methods for learning structures of probabilistic graphical models. In this chapter, we focus on the theoretical properties of the parametrization of graphical models.

We show that the log-likelihood of several probabilistic graphical models is Lipschitz continuous with respect to the ℓ_p-norm of the parameters. We discuss several implications of Lipschitz parametrization. We present an upper bound of the Kullback-Leibler divergence that allows understanding methods that penalize the ℓ_p-norm of differences of parameters as the minimization of that upper bound. The expected log-likelihood is lower bounded by the negative ℓ_p-norm, which allows understanding the generalization ability of probabilistic models. The exponential of the negative ℓ_p-norm is involved in the lower bound of the Bayes error rate, which shows that it is reasonable to use parameters as features in algorithms that rely on metric spaces (e.g. classification, dimensionality reduction, clustering). Our results do not rely on specific algorithms for learning the structure or parameters. We show preliminary results for activity recognition and temporal segmentation.

6.1 Introduction

Several methods have been proposed for learning the structure and parameters of graphical models from data. We mention only a few references that follow a maximum likelihood approach for Markov random fields [Lee et al., 2006a], Ising models [Höfling and Tibshirani, 2009], Gaussian graphical models [Banerjee et al., 2006, Friedman et al., 2007b] and Bayesian networks [Guo and Schuurmans, 2006, Schmidt et al., 2007b]. One may ask whether the log-likelihood is “well behaved”, i.e. small changes in the parameters produce small changes in the objective function. Another natural question is whether the ℓ_p distance between the learnt parameters and the ground truth provides some guarantee on their generalization ability, i.e. the expected log-likelihood.

When learning multiple graphical models, several authors have proposed ℓ_p-norm regularizers from the difference of parameters between two models. Zhang and Wang [2010] proposed a method that detects sparse structural changes of Gaussian graphical models in controlled experiments between two experimental conditions. Kolar et al. [2010] proposed a total variation regularizer for learning time-varying Ising models with sparse changes along the time course. Kolar et al. [2009] proposed a similar method for Gaussian graphical models. One natural question is whether the ℓ_p-norm of the difference of parameters between two graphical models is related to a measure of similarity between probability distributions, i.e. the Kullback-Leibler divergence.

There are several experimental results where the parameters of graphical models were used as features for classification and clustering. Classification of image textures from the precision matrix of Gaussian graphical models as features was proposed in [Chellappa and Chatterjee, 1985], and from parameters of Ising models in [Chen and Dubes, 1990]. The use of the covariance matrix as features for detection of humans in still images was proposed in [Tuzel et al., 2007]. Clustering by using the Gaussian graphical model parameters was performed in [Kolar et al., 2009], where they show discriminability between different type of imaginations from electroencephalography (EEG) recordings. One may ask whether the parameters of graphical models approximately lie in an metric space (ℓ_p) that allows for classification and clustering. In other words, whether the ℓ_p-norm of the difference of parameters between two graphical models is related to a measure of discriminability, i.e. the Bayes error rate.

In this chapter, we define Lipschitz continuous parametrization of probabilistic models. Through Lipschitz parametrization, we provide an upper bound of the Kullback-Leibler divergence. Therefore, methods that penalize the ℓ_p-norm of differences of parameters [Kolar et al., 2009, 2010, Zhang and Wang, 2010] are minimizing an upper bound of the Kullback-Leibler divergence. We show that Lipschitz parametrization also allows understanding the generalization ability of probabilistic models by providing a lower bound of the expected log-likelihood. Finally, we provide a lower bound of the Bayes error rate that depends on the ℓ_p-norm of the model parameters. This allows understanding the use of model parameters as features for classification and clustering as in [Chellappa and Chatterjee, 1985, Chen and Dubes, 1990, Tuzel et al., 2007, Kolar et al., 2009].

We believe that Lipschitz parametrization is a natural definition since it implies the Lipschitz continuity of the sample log-likelihood, which might prove very useful for maximum likelihood estimation. Furthermore, we show which parametrization is Lipschitz. For instance, in the case of discrete Bayesian networks, conditional probability tables are represented in an exponential space that resembles the *softmax activation function*.

6.2 Preliminaries

In this section, we introduce probabilistic graphical models and Lipschitz continuity.

We assume $\mathbf{x} \in \mathbb{R}^N$ for continuous random variables. For discrete random variables, we assume $\mathbf{x} \in \times_n \{1, \ldots, X_n\}$, i.e. $(\forall n)\ x_n \in \{1, \ldots, X_n\}$. First, we define three general classes of graphical models: *Bayesian networks*, *Markov random fields* and *factor graphs*. It is well known that factor graphs subsume Bayesian networks and Markov random fields. Our choice of analyzing all types of graphical models comes from the fact that we use different parametrizations for each type.

Definition 6.1. *A Bayesian network [Koller and Friedman, 2009, Lauritzen, 1996] for random variables* $\mathbf{x}$ *is a directed acyclic graph with one conditional probability function* $p(x_n|\mathbf{x}_{\pi_n})$ *for each variable* x_n *given its set of parents* $\pi_n \subseteq \{1, \ldots, N\}$*. The joint probability distribution is given by:*

$$p(\mathbf{x}) = \prod_n p(x_n|\mathbf{x}_{\pi_n}) \tag{6.1}$$

where $(\forall n, \mathbf{x}_{\pi_n})\ \int_{x_n} p(x_n|\mathbf{x}_{\pi_n}) = 1$ *and therefore* $p(\mathbf{x})$ *is valid, i.e.* $\int_{\mathbf{x}} p(\mathbf{x}) = 1$.

Definition 6.2. *A Markov random field [Koller and Friedman, 2009, Lauritzen, 1996] for random variables* $\mathbf{x}$ *is an undirected graph with one potential function* ϕ_c *for each maximal clique* $\varphi_c \subseteq \{1, \ldots, N\}$*. The joint probability distribution is given by:*

$$p(\mathbf{x}) = \frac{1}{\mathcal{Z}} \prod_c \phi_c(\mathbf{x}_{\varphi_c}) \tag{6.2}$$

where the partition function $\mathcal{Z} = \int_{\mathbf{x}} \prod_c \phi_c(\mathbf{x}_{\varphi_c})$ *ensures that* $p(\mathbf{x})$ *is valid, i.e.* $\int_{\mathbf{x}} p(\mathbf{x}) = 1$.

Definition 6.3. *A factor graph [Koller and Friedman, 2009] for random variables* $\mathbf{x}$ *is a bipartite graph where one set of nodes are the random variables and the other set are the local functions. Each local function* ϕ_c *is connected to the set variables of variables* $\varphi_c \subseteq \{1, \ldots, N\}$ *on which it depends on. The joint probability distribution is given by:*

$$p(\mathbf{x}) = \frac{1}{\mathcal{Z}} \prod_c \phi_c(\mathbf{x}_{\varphi_c}) \tag{6.3}$$

where the partition function $\mathcal{Z} = \int_{\mathbf{x}} \prod_c \phi_c(\mathbf{x}_{\varphi_c})$ *ensures that* $p(\mathbf{x})$ *is valid, i.e.* $\int_{\mathbf{x}} p(\mathbf{x}) = 1$.

For completeness, we introduce Lipschitz continuity for differentiable functions.

Definition 6.4. *Given the parameters* $\mathbf{\Theta} \in \mathbb{R}^{M_1 \times M_2}$*, a differentiable function* $f(\mathbf{\Theta}) \in \mathbb{R}$ *is called Lipschitz continuous with respect to the* ℓ_p*-norm of* $\mathbf{\Theta}$*, if there exists a constant* $K \geq 0$ *such that:*

$$(\forall \mathbf{\Theta}_1, \mathbf{\Theta}_2)\ |f(\mathbf{\Theta}_1) - f(\mathbf{\Theta}_2)| \leq K \|\mathbf{\Theta}_1 - \mathbf{\Theta}_2\|_p \tag{6.4}$$

or equivalently:

$$(\forall \mathbf{\Theta})\ \|\partial f / \partial \mathbf{\Theta}\|_p \leq K \tag{6.5}$$

6.3 Lipschitz Parametrization and Implications

In this section, we define Lipschitz parametrization of probabilistic models and discuss its implications.

6.3.1 Lipschitz Parametrization

We extend the Lipschitz continuity notion to the parametrization of probability distributions.

Definition 6.5. *A probability distribution $\mathcal{P} = p(\cdot|\mathbf{\Theta})$ parameterized by $\mathbf{\Theta} \in \mathbb{R}^{M_1 \times M_2}$ is called (ℓ_p, K)-Lipschitz continuous if for all $\mathbf{x}$, the log-likelihood $f(\mathbf{\Theta}) = \log p(\mathbf{x}|\mathbf{\Theta})$ is Lipschitz continuous with respect to the ℓ_p-norm of $\mathbf{\Theta}$ with constant $K(\mathbf{x})$.*

Remark 6.6. *Note that (ℓ_p, K)-Lipschitz continuity implies $(\ell_{p'}, K')$-Lipschitz continuity, since all vector and matrix norms are equivalent, i.e. $(\forall \mathbf{\Theta} \in \mathbb{R}^{M_1 \times M_2})$ $\alpha \|\mathbf{\Theta}\|_p \leq \|\mathbf{\Theta}\|_{p'} \leq \beta \|\mathbf{\Theta}\|_p$ for some $\alpha, \beta > 0$ and $M_1, M_2 < +\infty$.*

If we are interested in Euclidean spaces, we would need to prove Lipschitz continuity with respect to the ℓ_2-norm for vectors or the Frobenius norm for matrices. Due to Remark 6.6, we can chose any particular norm for proving Lipschitz continuity.

6.3.2 Kullback-Leibler Divergence

We show that the ℓ_p-norm is an upper bound of the Kullback-Leibler divergence. Therefore, methods that penalize the ℓ_p-norm of differences of parameters are minimizing an upper bound of the Kullback-Leibler divergence.

We chose Kullback-Leibler divergence for being one of the most used, and because it includes the "log p" term, which relates to our Lipschitz continuity definition. For this reason, it is straightforward to derive similar bounds for the Jensen-Shannon and Jeffrey's divergences. Additionally, there are several lower bounds of the Kullback-Leibler divergence (e.g. variational distance and Hellinger's distance). Therefore, our upperbound on Kullback-Leibler also upperbounds these other divergence measures.

Theorem 6.7. *Given two (ℓ_p, K)-Lipschitz continuous distributions $\mathcal{P}_1 = p(\cdot|\mathbf{\Theta}_1)$ and $\mathcal{P}_2 = p(\cdot|\mathbf{\Theta}_2)$, the Kullback-Leibler divergence from $\mathcal{P}_1$ to $\mathcal{P}_2$ is bounded as follows:*

$$\mathcal{KL}(\mathcal{P}_1||\mathcal{P}_2) \leq \overline{K} \|\mathbf{\Theta}_1 - \mathbf{\Theta}_2\|_p \tag{6.6}$$

with constant $\overline{K} = \mathbb{E}_{\mathcal{P}_1}[K(\mathbf{x})]$.

Proof. By definition $\mathcal{KL}(\mathcal{P}_1||\mathcal{P}_2) = \mathbb{E}_{\mathcal{P}_1}[\log p(\mathbf{x}|\mathbf{\Theta}_1) - \log p(\mathbf{x}|\mathbf{\Theta}_2)] \leq \mathbb{E}_{\mathcal{P}_1}[|\log p(\mathbf{x}|\mathbf{\Theta}_1) - \log p(\mathbf{x}|\mathbf{\Theta}_2)|] \equiv B$. Note that by Definitions 6.4 and 6.5, $B \leq \mathbb{E}_{\mathcal{P}_1}[K(\mathbf{x}) \|\mathbf{\Theta}_1 - \mathbf{\Theta}_2\|_p] = \mathbb{E}_{\mathcal{P}_1}[K(\mathbf{x})] \|\mathbf{\Theta}_1 - \mathbf{\Theta}_2\|_p = \overline{K} \|\mathbf{\Theta}_1 - \mathbf{\Theta}_2\|_p$. □

Remark 6.8. *For identifiable distributions $\mathcal{P}_1 = p(\cdot|\mathbf{\Theta}_1)$ and $\mathcal{P}_2 = p(\cdot|\mathbf{\Theta}_2)$ (i.e. $\mathcal{P}_1 = \mathcal{P}_2 \Rightarrow \mathbf{\Theta}_1 = \mathbf{\Theta}_2$), the upper bound in Theorem 6.7 is tight since the Kullback-Leibler divergence is zero if and only if the parameters are equal. More formally, $\mathcal{KL}(\mathcal{P}_1||\mathcal{P}_2) = 0 \Leftrightarrow \mathcal{P}_1 = \mathcal{P}_2 \Leftrightarrow \mathbf{\Theta}_1 = \mathbf{\Theta}_2 \Leftrightarrow \|\mathbf{\Theta}_1 - \mathbf{\Theta}_2\|_p = 0$.*

Remark 6.9. *The upper bound in Theorem 6.7 also applies for every marginal distribution by properties of the Kullback-Leibler divergence.*

6.3.3 Expected Log-Likelihood

We show the importance of Lipschitz continuity for understanding the generalization ability of probabilistic models, by showing that the negative ℓ_p-norm is a lower bound of the expected log-likelihood.

Theorem 6.10. *Given two (ℓ_p, K)-Lipschitz continuous distributions $\mathcal{P} = p(\cdot|\mathbf{\Theta})$ and $\mathcal{P}^* = p(\cdot|\mathbf{\Theta}^*)$, the expected log-likelihood (also called negative cross entropy) of the learnt distribution $\mathcal{P}$ with respect to the ground truth distribution $\mathcal{P}^*$ is bounded as follows:*

$$-\mathcal{H}(\mathcal{P}^*) - \overline{K}\|\mathbf{\Theta}^* - \mathbf{\Theta}\|_p \leq \mathbb{E}_{\mathcal{P}^*}[\log p(\mathbf{x}|\mathbf{\Theta})] \leq 0 \tag{6.7}$$

with constant $\overline{K} = \mathbb{E}_{\mathcal{P}^*}[K(\mathbf{x})]$.

Proof. Since $0 = -\mathcal{H}(\mathcal{P}^*) - \mathbb{E}_{\mathcal{P}^*}[\log p(\mathbf{x}|\mathbf{\Theta}^*)]$ we have $\mathbb{E}_{\mathcal{P}^*}[\log p(\mathbf{x}|\mathbf{\Theta})] = \mathbb{E}_{\mathcal{P}^*}[\log p(\mathbf{x}|\mathbf{\Theta})] - \mathcal{H}(\mathcal{P}^*) - \mathbb{E}_{\mathcal{P}^*}[\log p(\mathbf{x}|\mathbf{\Theta}^*)] = -\mathcal{H}(\mathcal{P}^*) - \mathbb{E}_{\mathcal{P}^*}[\log p(\mathbf{x}|\mathbf{\Theta}^*) - \log p(\mathbf{x}|\mathbf{\Theta})] = -\mathcal{H}(\mathcal{P}^*) - \mathcal{KL}(\mathcal{P}^*||\mathcal{P})$. The upper bound follows from the non-negativity of the Kullback-Leibler divergence and entropy.

For proving the lower bound, given that $\mathcal{KL}(\mathcal{P}^*||\mathcal{P}) \leq \overline{K}\|\mathbf{\Theta}^* - \mathbf{\Theta}\|_p$ by Theorem 6.7, we prove our claim. □

In the following Section 6.4, we prove that for probabilistic models over discrete random variables, $(\forall \mathbf{x})\ K(\mathbf{x}) = 1$ and therefore $\overline{K} = 1$. For continuous random variables, given its generality, the constant $K(\mathbf{x})$ depends on $\mathbf{x}$ and therefore $\overline{K}$ is looser and does not have a closed-form expression; except for specific cases, e.g. Gaussian graphical models.

6.3.4 Bayes Error Rate

We show the importance of Lipschitz continuity for discriminability, by showing that the exponential of the negative ℓ_p-norm is involved in lower bound of the Bayes error rate. This allows understanding the use of model parameters as features for classification and clustering. We also motivate a distance measure similar to the Chernoff bound [Chernoff, 1952], i.e. the negative log-Bayes error rate.

In the next theorem, given two classes ϖ_1 and ϖ_2 we assumed priors $P(\varpi_1) = P(\varpi_2) = \frac{1}{2}$ for clarity of presentation. It is straightforward to state a more general result for arbitrary $P(\varpi_1) + P(\varpi_2) = 1$.

Theorem 6.11. *Given two classes ϖ_1 and ϖ_2 with priors $P(\varpi_1) = P(\varpi_2) = \frac{1}{2}$ and their corresponding (ℓ_p, K)-Lipschitz continuous distributions $\mathcal{P}_1 = p(\cdot|\mathbf{\Theta}_1)$ and $\mathcal{P}_2 = p(\cdot|\mathbf{\Theta}_2)$, the Bayes error rate $\mathcal{BE}(\mathbf{\Theta}_1, \mathbf{\Theta}_2) = \frac{1}{2}\int_{\mathbf{x}} \min(p(\mathbf{x}|\mathbf{\Theta}_1), p(\mathbf{x}|\mathbf{\Theta}_2))$ is bounded as follows:*

$$\frac{\mathcal{BB}(\mathbf{\Theta}_1, \mathbf{\Theta}_2)}{4} \leq \mathcal{BE}(\mathbf{\Theta}_1, \mathbf{\Theta}_2) \tag{6.8}$$

$$\log 2 \leq -\log \mathcal{BE}(\mathbf{\Theta}_1, \mathbf{\Theta}_2) \leq \log 4 + \widetilde{K}\|\mathbf{\Theta}_1 - \mathbf{\Theta}_2\|_p \tag{6.9}$$

where $\mathcal{BB}(\mathbf{\Theta}_1, \mathbf{\Theta}_2) = \sum_c \mathbb{E}_{\mathcal{P}_c}[e^{-K(\mathbf{x})\|\mathbf{\Theta}_1 - \mathbf{\Theta}_2\|_p}]$ and $\widetilde{K} = \min_c \mathbb{E}_{\mathcal{P}_c}[K(\mathbf{x})]$.

Proof. Let $p_c \equiv p(\mathbf{x}|\mathbf{\Theta}_c)$. We can rewrite $\mathcal{BE}(\mathbf{\Theta}_1, \mathbf{\Theta}_2) = \frac{1}{2}\int_{\mathbf{x}} \min\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right)(p_1 + p_2) = \frac{1}{2}\int_{\mathbf{x}} e^{\min\left(\log \frac{p_1}{p_1+p_2}, \log \frac{p_2}{p_1+p_2}\right)}(p_1 + p_2)$. We can also rewrite $\log \frac{p_1}{p_1+p_2} = -\log\left(1 + \frac{p_2}{p_1}\right) = -\ell(z_{12})$, where $z_{12} = \log p_1 - \log p_2$ and $\ell(z) = \log(1 + e^{-z})$ is the logistic loss. Similarly $\log \frac{p_2}{p_1+p_2} = -\ell(-z_{12})$. Therefore $\min\left(\log \frac{p_1}{p_1+p_2}, \log \frac{p_2}{p_1+p_2}\right) = \min(-\ell(z_{12}), -\ell(-z_{12}))$. Note that $(\forall z)$ −

$|z| - \log 2 \leq \min(-\ell(z), -\ell(-z))$. Since both $\mathcal{P}_1$ and $\mathcal{P}_2$ are (ℓ_p, K)-Lipschitz continuous, by Definitions 6.4 and 6.5, we have $-K(\mathbf{x})\|\mathbf{\Theta}_1 - \mathbf{\Theta}_2\|_p - \log 2 \leq \min\left(\log \frac{p_1}{p_1+p_2}, \log \frac{p_2}{p_1+p_2}\right)$.

For proving the lower bound in eq.(6.8), $\mathcal{BE}(\mathbf{\Theta}_1, \mathbf{\Theta}_2) \geq \frac{1}{2} \int_{\mathbf{x}} e^{-K(\mathbf{x})\|\mathbf{\Theta}_1 - \mathbf{\Theta}_2\|_p - \log 2}(p_1 + p_2) = \frac{1}{4} \int_{\mathbf{x}} e^{-K(\mathbf{x})\|\mathbf{\Theta}_1 - \mathbf{\Theta}_2\|_p}(p_1 + p_2) = \frac{1}{4}\mathcal{BB}(\mathbf{\Theta}_1, \mathbf{\Theta}_2)$.

The lower bound of eq.(6.9) follows from the fact that $\mathcal{BE}(\mathbf{\Theta}_1, \mathbf{\Theta}_2) \leq \frac{1}{2}$. For proving the upper bound of eq.(6.9), by Jensen's inequality $\frac{1}{4} \sum_c e^{-\mathbb{E}_{\mathcal{P}_c}[K(\mathbf{x})]\|\mathbf{\Theta}_1 - \mathbf{\Theta}_2\|_p} \leq \frac{\mathcal{BB}(\mathbf{\Theta}_1, \mathbf{\Theta}_2)}{4} \leq \mathcal{BE}(\mathbf{\Theta}_1, \mathbf{\Theta}_2)$. Therefore, $-\log \mathcal{BE}(\mathbf{\Theta}_1, \mathbf{\Theta}_2) \leq \log 4 - \log \sum_c e^{-\mathbb{E}_{\mathcal{P}_c}[K(\mathbf{x})]\|\mathbf{\Theta}_1 - \mathbf{\Theta}_2\|_p}$. By properties of the logsumexp function, we have $-\log \mathcal{BE}(\mathbf{\Theta}_1, \mathbf{\Theta}_2) \leq \log 4 - \max_c\left(-\mathbb{E}_{\mathcal{P}_c}[K(\mathbf{x})]\right)\|\mathbf{\Theta}_1 - \mathbf{\Theta}_2\|_p = \log 4 + \min_c \mathbb{E}_{\mathcal{P}_c}[K(\mathbf{x})]\|\mathbf{\Theta}_1 - \mathbf{\Theta}_2\|_p$. □

6.4 Lipschitz Continuous Models

In this section, we show that several probabilistic graphical models are Lipschitz continuous. This includes Bayesian networks, Markov random fields and factor graphs for discrete and continuous random variables. Dynamic models such as dynamic Bayesian networks and conditional random fields are also Lipschitz continuous.

6.4.1 Bayesian Networks

We show that a sufficient condition for the Lipschitz continuity of Bayesian networks is the Lipschitz continuity of the conditional probability functions.

Lemma 6.12. *For each variable x_n, given a (ℓ_p, K)-Lipschitz continuous conditional probability function $p(x_n|\mathbf{x}_{\pi_n}, \mathbf{\Theta})$, the Bayesian network $p(\mathbf{x}|\mathbf{\Theta}) = \prod_n p(x_n|\mathbf{x}_{\pi_n}, \mathbf{\Theta})$ is (ℓ_p, NK)-Lipschitz continuous.*

Proof. Let $g_n(\mathbf{\Theta}) = \log p(x_n|\mathbf{x}_{\pi_n}, \mathbf{\Theta})$ and $f(\mathbf{\Theta}) = \log p(\mathbf{x}|\mathbf{\Theta}) = \sum_n \log p(x_n|\mathbf{x}_{\pi_n}, \mathbf{\Theta}) = \sum_n g_n(\mathbf{\Theta})$, and therefore $\partial f/\partial\mathbf{\Theta} = \sum_n \partial g_n/\partial\mathbf{\Theta}$. By Definitions 6.4 and 6.5, we have $\|\partial f/\partial\mathbf{\Theta}\|_p \leq \sum_n \|\partial g_n/\partial\mathbf{\Theta}\|_p \leq NK(\mathbf{x})$. □

Remark 6.13. *When comparing two Bayesian networks, the set of parents π_n for each variable x_n is not necessarily the same for both networks. Since Lemma 6.12 does not use the fact that the joint probability distribution $p(\mathbf{x}|\mathbf{\Theta})$ is valid (i.e. $\int_{\mathbf{x}} p(\mathbf{x}|\mathbf{\Theta}) = 1$ which is given by the acyclicity constraints), we can join the set of parents of both Bayesian networks before comparing them. More formally, let $\pi_n^{(1)}$ and $\pi_n^{(2)}$ be the set of parents of variable x_n in Bayesian network 1 and 2 respectively. It is trivial to show that if $p(x_n|\mathbf{x}_{\pi_n^{(1)}}, \mathbf{\Theta})$ and $p(x_n|\mathbf{x}_{\pi_n^{(2)}}, \mathbf{\Theta})$ are Lipschitz continuous, so is $p(x_n|\mathbf{x}_{\pi_n^{(1)} \cup \pi_n^{(2)}}, \mathbf{\Theta})$.*

Given the previous discussion, in the sequel, we show Lipschitz continuity for the conditional probability functions only.

6.4.2 Discrete Bayesian Networks

The following parametrization of Bayesian networks for discrete random variables is equivalent to using conditional probability tables. We use a representation in an exponential space that resembles the *softmax activation function* in the neural networks literature [Duda et al., 2001].

Lemma 6.14. *Let x_{π_n} be one of the possible parent value combinations for variable x_n, i.e. $x_{\pi_n} \in \{1, \dots, X_{\pi_n}\}$ where $X_{\pi_n} = \prod_{n' \in \pi_n} X_{n'}$. The conditional probability mass function for the discrete Bayesian network parameterized by $\mathbf{\Theta} = \{\mathbf{w}^{(n,1)}, \dots, \mathbf{w}^{(n,X_{\pi_n})}\}_n$, $(\forall n, x_{\pi_n})$ $\mathbf{w}^{(n,x_{\pi_n})} \in \mathbb{R}^{X_n - 1}$:*

$$\mathbb{P}[x_n = i | x_{\pi_n} = j, \mathbf{\Theta}] = \frac{e^{w_i^{(n,j)} 1[i < X_n]}}{\sum_{x_n} e^{w_{x_n}^{(n,j)}} + 1} \tag{6.10}$$

is $(\ell_\infty, 1)$-Lipschitz continuous.

Proof. Let $\mathbf{w} \equiv \mathbf{w}^{(n,x_{\pi_n})}$. For $i < X_n$, let $f(\mathbf{w}) = \log \mathbb{P}[x_n = i | x_{\pi_n} = j, \mathbf{\Theta}] = w_i - \log(\sum_{x_n} e^{w_{x_n}} + 1)$. By deriving $\partial f / \partial w_i = 1 - \frac{e^{w_i}}{\sum_{x_n} e^{w_{x_n}} + 1} = 1 - \mathbb{P}[x_n = i | x_{\pi_n} = j, \mathbf{\Theta}]$. Since $(\forall i)$ $0 \leq \mathbb{P}[x_n = i | x_{\pi_n} = j, \mathbf{\Theta}] \leq 1$, it follows that $|\partial f / \partial w_i| \leq 1$ and therefore $\|\partial f / \partial \mathbf{w}\|_\infty \leq 1$. By Definitions 6.4 and 6.5, we prove our claim. □

The following parametrization of Bayesian networks for discrete random variables corresponds to the *multinomial logistic regression.* It reduces to logistic regression for binary variables.

Lemma 6.15. *Given a feature function with F features $\boldsymbol{\psi}(\mathbf{x}_{\pi_n}) = (\psi_1(\mathbf{x}_{\pi_n}), \dots, \psi_F(\mathbf{x}_{\pi_n}))^{\mathrm{T}}$ such that $(\forall \mathbf{x}_{\pi_n})$ $\|\boldsymbol{\psi}(\mathbf{x}_{\pi_n})\|_\infty \leq 1$, the conditional probability mass function for the discrete Bayesian network parameterized by $\mathbf{\Theta} = \{\mathbf{w}_{(1)}^{(n)}, \dots, \mathbf{w}_{(X_n - 1)}^{(n)}\}_n$, $(\forall n, x_n)$ $\mathbf{w}_{(x_n)}^{(n)} \in \mathbb{R}^F$:*

$$\mathbb{P}[x_n = i | \mathbf{x}_{\pi_n}, \mathbf{\Theta}] = \frac{e^{\mathbf{w}_{(i)}^{(n)\mathrm{T}} \boldsymbol{\psi}(\mathbf{x}_{\pi_n}) 1[i < X_n]}}{\sum_{x_n} e^{\mathbf{w}_{(x_n)}^{(n)}{}^{\mathrm{T}} \boldsymbol{\psi}(\mathbf{x}_{\pi_n})} + 1} \tag{6.11}$$

is $(\ell_\infty, 1)$-Lipschitz continuous.

Proof. Let $\mathbf{w} \equiv \mathbf{w}^{(n)}$. For $i < X_n$, let $f(\mathbf{w}) = \mathbb{P}[x_n = i | \mathbf{x}_{\pi_n}, \mathbf{\Theta}] = \mathbf{w}_{(i)}{}^{\mathrm{T}} \boldsymbol{\psi}(\mathbf{x}_{\pi_n}) - \log(\sum_{x_n} e^{\mathbf{w}_{(x_n)}{}^{\mathrm{T}} \boldsymbol{\psi}(\mathbf{x}_{\pi_n})} + 1)$. By deriving $\partial f / \partial \mathbf{w}_{(i)} = \boldsymbol{\psi}(\mathbf{x}_{\pi_n}) - \frac{e^{\mathbf{w}_{(i)}{}^{\mathrm{T}} \boldsymbol{\psi}(\mathbf{x}_{\pi_n})} \boldsymbol{\psi}(\mathbf{x}_{\pi_n})}{\sum_{x_n} e^{\mathbf{w}_{(x_n)}{}^{\mathrm{T}} \boldsymbol{\psi}(\mathbf{x}_{\pi_n})} + 1} = (1 - \mathbb{P}[x_n = i | \mathbf{x}_{\pi_n}, \mathbf{\Theta}]) \boldsymbol{\psi}(\mathbf{x}_{\pi_n})$. Since $(\forall i)$ $0 \leq \mathbb{P}[x_n = i | \mathbf{x}_{\pi_n}, \mathbf{\Theta}] \leq 1$, it follows that $\|\partial f / \partial \mathbf{w}_{(i)}\|_\infty \leq \|\boldsymbol{\psi}(\mathbf{x}_{\pi_n})\|_\infty \leq 1$. By Definitions 6.4 and 6.5, we prove our claim. □

Note that the requirement that $(\forall \mathbf{x})$ $\|\boldsymbol{\psi}(\mathbf{x}_{\pi_n})\|_\infty \leq 1$ is not restrictive, since the random variables are discrete and we can perform scaling of the features.

6.4.3 Continuous Bayesian Networks

We focus on two types of continuous random variables: Gaussian and Laplace. For the Gaussian Bayesian network, we assume that the weight vector $\mathbf{w}$ of linear regression has bounded norm, i.e. $\|\mathbf{w}\|_2 \leq \beta$ (please, see Appendix G). We also assume that the features are normalized, i.e. the standard deviation is one.

Lemma 6.16. *Given a feature function with F features $\boldsymbol{\psi}(\mathbf{x}_{\pi_n}) = (\psi_1(\mathbf{x}_{\pi_n}), \ldots, \psi_F(\mathbf{x}_{\pi_n}))^{\mathrm{T}}$, the conditional probability density function for the Gaussian Bayesian network parameterized by $\boldsymbol{\Theta} = \{\mathbf{w}^{(n)}\}_n$, $(\forall n)$ $\mathbf{w}^{(n)} \in \mathbb{R}^F$:*

$$p(x_n|\mathbf{x}_{\pi_n}, \boldsymbol{\Theta}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_n - \mathbf{w}^{(n)\mathrm{T}}\boldsymbol{\psi}(\mathbf{x}_{\pi_n}))^2} \tag{6.12}$$

is $(\ell_2, \|\boldsymbol{\psi}(\mathbf{x}_{\pi_n})\|_2|x_n| + \beta\|\boldsymbol{\psi}(\mathbf{x}_{\pi_n})\|_2^2)$-Lipschitz continuous.

Proof. Let $\mathbf{w} \equiv \mathbf{w}^{(n)}$ and $f(\mathbf{w}) = \log p(x_n|\mathbf{x}_{\pi_n}, \boldsymbol{\Theta}) = \frac{1}{2}(-\log(2\pi) - (x_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{x}_{\pi_n}))^2)$. By deriving $\partial f/\partial \mathbf{w} = (x_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{x}_{\pi_n}))\boldsymbol{\psi}(\mathbf{x}_{\pi_n})$. Then $\|\partial f/\partial \mathbf{w}\|_2 \leq |x_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{x}_{\pi_n})| \; \|\boldsymbol{\psi}(\mathbf{x}_{\pi_n})\|_2 \leq (|x_n| + |\mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{x}_{\pi_n})|) \; \|\boldsymbol{\psi}(\mathbf{x}_{\pi_n})\|_2 \leq (|x_n| + \|\mathbf{w}\|_2\|\boldsymbol{\psi}(\mathbf{x}_{\pi_n})\|_2) \; \|\boldsymbol{\psi}(\mathbf{x}_{\pi_n})\|_2$. By noting that $\|\mathbf{w}\|_2 \leq \beta$ and by Definitions 6.4 and 6.5, we prove our claim. □

Remark 6.17. *In Lemma 6.16, the expression $K(\mathbf{x}) = \|\boldsymbol{\psi}(\mathbf{x}_{\pi_n})\|_2|x_n| + \beta\|\boldsymbol{\psi}(\mathbf{x}_{\pi_n})\|_2^2$ becomes more familiar for a linear feature function $\boldsymbol{\psi}(\mathbf{x}_{\pi_n}) = \mathbf{x}_{\pi_n}$. In this case, note that $(\forall \pi_n)$ $\|\boldsymbol{\psi}(\mathbf{x}_{\pi_n})\|_2 = \|\mathbf{x}_{\pi_n}\|_2 \leq \|\mathbf{x}\|_2$ and $(\forall n)$ $|x_n| \leq \|\mathbf{x}\|_2$. Therefore $K(\mathbf{x}) \leq (1+\beta)\|\mathbf{x}\|_2^2$.*

For the Laplace Bayesian network, we assume that the features are normalized, i.e. the absolute deviation is one.

Lemma 6.18. *Given a feature function with F features $\boldsymbol{\psi}(\mathbf{x}_{\pi_n}) = (\psi_1(\mathbf{x}_{\pi_n}), \ldots, \psi_F(\mathbf{x}_{\pi_n}))^{\mathrm{T}}$, the conditional probability density function for the Laplace Bayesian network parameterized by $\boldsymbol{\Theta} = \{\mathbf{w}^{(n)}\}_n$, $(\forall n)$ $\mathbf{w}^{(n)} \in \mathbb{R}^F$:*

$$p(x_n|\mathbf{x}_{\pi_n}, \boldsymbol{\Theta}) = \frac{1}{2} e^{-|x_n - \mathbf{w}^{(n)\mathrm{T}}\boldsymbol{\psi}(\mathbf{x}_{\pi_n})|} \tag{6.13}$$

is $(\ell_2, \|\boldsymbol{\psi}(\mathbf{x}_{\pi_n})\|_2)$-Lipschitz continuous.

Proof. Let $\mathbf{w} \equiv \mathbf{w}^{(n)}$ and $f(\mathbf{w}) = \log p(x_n|\mathbf{x}_{\pi_n}, \boldsymbol{\Theta}) = -\log 2 - |x_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{x}_{\pi_n})|$. The subdifferential set of the non-smooth function f can be written as $\partial f/\partial \mathbf{w} = \boldsymbol{\psi}(\mathbf{x}_{\pi_n})s(x_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{x}_{\pi_n}))$, where $s(z) = +1$ for $z > 0$, $s(z) = -1$ for $z < 0$ and $s(z) \in [-1; +1]$ for $z = 0$. Therefore $\|\partial f/\partial \mathbf{w}\|_2 \leq \|\boldsymbol{\psi}(\mathbf{x}_{\pi_n})\|_2$. By Definitions 6.4 and 6.5, we prove our claim. □

6.4.4 Discrete Factor Graphs

The following parameterization of factor graphs for discrete random variables includes Markov random fields when the features depend on the cliques. A special case of this parametrization are Ising models (i.e. Markov random fields on binary variables with pairwise interactions). The feature function $\boldsymbol{\psi}(\mathbf{x}) = (\mathbf{vec}(\mathbf{x}\mathbf{x}^{\mathrm{T}}), \mathbf{x})$ for Ising models with external field, and $\boldsymbol{\psi}(\mathbf{x}) = \mathbf{vec}(\mathbf{x}\mathbf{x}^{\mathrm{T}})$ without external field.

Lemma 6.19. *Given a feature function with F features $\boldsymbol{\psi}(\mathbf{x}) = (\psi_1(\mathbf{x}), ..., \psi_F(\mathbf{x}))^{\mathrm{T}}$ such that $(\forall \mathbf{x})$ $\|\boldsymbol{\psi}(\mathbf{x})\|_\infty \leq 1$, the discrete factor graph $\mathcal{P} = p(\cdot|\boldsymbol{\Theta})$ parameterized by $\boldsymbol{\Theta} = \mathbf{w}$, $\mathbf{w} \in \mathbb{R}^F$ with probability mass function:*

$$p(\mathbf{x}|\boldsymbol{\Theta}) = \frac{1}{\mathcal{Z}(\mathbf{w})} e^{\mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{x})} \tag{6.14}$$

where $\mathcal{Z}(\mathbf{w}) = \sum_{\mathbf{x}} e^{\mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{x})}$ is $(\ell_\infty, 2)$-Lipschitz continuous.

Proof. Let $f(\mathbf{w}) = \log p(\mathbf{x}|\boldsymbol{\Theta}) = \mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{x}) - \log(\sum_{\mathbf{x}} e^{\mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{x})})$. By deriving $\partial f/\partial\mathbf{w} = \boldsymbol{\psi}(\mathbf{x}) - \frac{\sum_{\mathbf{x}} e^{\mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{x})}\boldsymbol{\psi}(\mathbf{x})}{\sum_{\mathbf{x}} e^{\mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{x})}} = \boldsymbol{\psi}(\mathbf{x}) - \mathbb{E}_{\mathcal{P}}[\boldsymbol{\psi}(\mathbf{x})]$. Since the expected value for discrete random variables is a weighted sum with positive weights that add up to 1 and $(\forall\mathbf{x})\ \|\boldsymbol{\psi}(\mathbf{x})\|_\infty \leq 1$ therefore $\|\mathbb{E}_{\mathcal{P}}[\boldsymbol{\psi}(\mathbf{x})]\|_\infty \leq 1$. It follows that $\|\partial f/\partial\mathbf{w}\|_\infty \leq \|\boldsymbol{\psi}(\mathbf{x})\|_\infty + \|\mathbb{E}_{\mathcal{P}}[\boldsymbol{\psi}(\mathbf{x})]\|_\infty \leq 2$. By Definitions 6.4 and 6.5, we prove our claim. □

Note that the requirement that $(\forall\mathbf{x})\ \|\boldsymbol{\psi}(\mathbf{x})\|_\infty \leq 1$ is not restrictive, since the random variables are discrete and we can perform scaling of the features.

6.4.5 Continuous Factor Graphs

The following parameterization of factor graphs for continuous random variables includes Markov random fields when the features depend on the cliques. A special case of this parametrization are Gaussian graphical models (i.e. Markov random fields on jointly Gaussian variables), in which the feature function $\boldsymbol{\psi}(\mathbf{x}) = \mathbf{vec}(\mathbf{x}\mathbf{x}^{\mathrm{T}})$ and $\boldsymbol{\Theta} \succ \mathbf{0}$.

Lemma 6.20. *Given a feature function with F features $\boldsymbol{\psi}(\mathbf{x}) = (\psi_1(\mathbf{x}), ..., \psi_F(\mathbf{x}))^{\mathrm{T}}$ such that $\mathbb{E}_{\mathcal{P}}[\|\boldsymbol{\psi}(\mathbf{x})\|_p] \leq \alpha$, the continuous factor graph $\mathcal{P} = p(\cdot|\boldsymbol{\Theta})$ parameterized by $\boldsymbol{\Theta} = \mathbf{w}$, $\mathbf{w} \in \mathbb{R}^F$ with probability density function:*

$$p(\mathbf{x}|\boldsymbol{\Theta}) = \frac{1}{\mathcal{Z}(\mathbf{w})} e^{\mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{x})} \tag{6.15}$$

where $\mathcal{Z}(\mathbf{w}) = \int_{\mathbf{x}} e^{\mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{x})}$ is $(\ell_p, \|\boldsymbol{\psi}(\mathbf{x})\|_p + \alpha)$-Lipschitz continuous.

Proof. Let $f(\mathbf{w}) = \log p(\mathbf{x}|\boldsymbol{\Theta}) = \mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{x}) - \log(\int_{\mathbf{x}} e^{\mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{x})})$. By deriving $\partial f/\partial\mathbf{w} = \boldsymbol{\psi}(\mathbf{x}) - \frac{\int_{\mathbf{x}} e^{\mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{x})}\boldsymbol{\psi}(\mathbf{x})}{\int_{\mathbf{x}} e^{\mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{x})}} = \boldsymbol{\psi}(\mathbf{x}) - \mathbb{E}_{\mathcal{P}}[\boldsymbol{\psi}(\mathbf{x})]$. By Jensen's inequality $\|\mathbb{E}_{\mathcal{P}}[\boldsymbol{\psi}(\mathbf{x})]\|_p \leq \mathbb{E}_{\mathcal{P}}[\|\boldsymbol{\psi}(\mathbf{x})\|_p] \leq \alpha$. It follows that $\|\partial f/\partial\mathbf{w}\|_p \leq \|\boldsymbol{\psi}(\mathbf{x})\|_p + \|\mathbb{E}_{\mathcal{P}}[\boldsymbol{\psi}(\mathbf{x})]\|_p \leq \|\boldsymbol{\psi}(\mathbf{x})\|_p + \alpha$. By Definitions 6.4 and 6.5, we prove our claim. □

The requirement that $\mathbb{E}_{\mathcal{P}}[\|\boldsymbol{\psi}(\mathbf{x})\|_p] \leq \alpha$ is also useful in deriving a close-form expresion of the Kullback-Leibler divergence bound.

Lemma 6.21. *Given two continuous factor graphs as in eq.*(6.15)*, i.e. $\mathcal{P}_1 = p(\cdot|\boldsymbol{\Theta}_1)$ and $\mathcal{P}_2 = p(\cdot|\boldsymbol{\Theta}_2)$, the Kullback-Leibler divergence from $\mathcal{P}_1$ to $\mathcal{P}_2$ is bounded as follows:*

$$\mathcal{KL}(\mathcal{P}_1||\mathcal{P}_2) \leq 2\alpha\|\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2\|_p \tag{6.16}$$

Proof. By invoking Theorem 6.7, the Lipschitz constant $\overline{K} = \mathbb{E}_{\mathcal{P}_1}[K(\mathbf{x})]$. By invoking Lemma 6.20, $K(\mathbf{x}) = \|\boldsymbol{\psi}(\mathbf{x})\|_p + \alpha$ and $\mathbb{E}_{\mathcal{P}_1}[\|\boldsymbol{\psi}(\mathbf{x})\|_p] \leq \alpha$. Finally, $\mathbb{E}_{\mathcal{P}_1}[K(\mathbf{x})] = \mathbb{E}_{\mathcal{P}_1}[\|\boldsymbol{\psi}(\mathbf{x})\|_p] + \alpha \leq 2\alpha$. □

6.4.6 Gaussian Graphical Models

A *Gaussian graphical model* [Lauritzen, 1996] is a Markov random field in which all random variables are continuous and jointly Gaussian. This model corresponds to the multivariate normal distribution.

We first analyze parametrization by using precision matrices. This parametrization is natural since it corresponds to factors graphs as in eq.(6.15) and therefore conditional independence corresponds to zeros in the precision matrix. We assume that the precision matrix $\mathbf{\Omega}$ has bounded norm, i.e. $\alpha\mathbf{I} \preceq \mathbf{\Omega} \preceq \beta\mathbf{I}$ or equivalently $\|\mathbf{\Omega}^{-1}\|_2 \leq \frac{1}{\alpha}$ and $\|\mathbf{\Omega}\|_2 \leq \beta$. This condition holds for Tikhonov regularization as well as for sparseness promoting (ℓ_1) methods (please, see Appendix H).

Lemma 6.22. *Given the precision matrix $\mathbf{\Omega} \succ \mathbf{0}$, the Gaussian graphical model parameterized by $\mathbf{\Theta} = \mathbf{\Omega}$, $\mathbf{\Omega} \in \mathbb{R}^{N\times N}$ with probability density function:*

$$p(\mathbf{x}|\mathbf{\Theta}) = \frac{(\det \mathbf{\Omega})^{1/2}}{(2\pi)^{N/2}} e^{-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{\Omega}\mathbf{x}} \tag{6.17}$$

is $(\ell_2, \frac{\|\mathbf{x}\|_2^2}{2} + \frac{1}{2\alpha})$-Lipschitz continuous.

Proof. Let $f(\mathbf{\Omega}) = \log p(\mathbf{x}|\mathbf{\Theta}) = \frac{1}{2}(\log\det\mathbf{\Omega} - N\log(2\pi) - \mathbf{x}^{\mathrm{T}}\mathbf{\Omega}\mathbf{x})$. By deriving $\partial f/\partial\mathbf{\Omega} = \frac{1}{2}(\mathbf{\Omega}^{-1} - \mathbf{x}\mathbf{x}^{\mathrm{T}})$. Therefore $\|\partial f/\partial\mathbf{\Omega}\|_2 \leq \frac{1}{2}(\|\mathbf{\Omega}^{-1}\|_2 + \|\mathbf{x}\mathbf{x}^{\mathrm{T}}\|_2) = \frac{1}{2}(\|\mathbf{\Omega}^{-1}\|_2 + \|\mathbf{x}\|_2^2) \leq \frac{1}{2}(\frac{1}{\alpha} + \|\mathbf{x}\|_2^2)$. By Definitions 6.4 and 6.5, we prove our claim. □

If we use Lemma 6.21, we will obtain a very loose bound of the Kullback-Leibler divergence where the constant $\overline{K} = \frac{2N\beta^{N/2}}{\alpha^{N/2+1}}$ (please, see Appendix I). Therefore, we analyze the specific case of Gaussian graphical models.

Lemma 6.23. *Given two Gaussian graphical models parameterized by their precision matrices as in eq.*(6.17)*, i.e. $\mathcal{P}_1 = p(\cdot|\mathbf{\Omega}_1)$ and $\mathcal{P}_2 = p(\cdot|\mathbf{\Omega}_2)$, the Kullback-Leibler divergence from $\mathcal{P}_1$ to $\mathcal{P}_2$:*

$$\mathcal{KL}(\mathcal{P}_1||\mathcal{P}_2) = \frac{1}{2}\left(\log\frac{\det\mathbf{\Omega}_1}{\det\mathbf{\Omega}_2} + \langle\mathbf{\Omega}_1^{-1}, \mathbf{\Omega}_2\rangle - N\right) \tag{6.18}$$

is bounded as follows:

$$\mathcal{KL}(\mathcal{P}_1||\mathcal{P}_2) \leq \frac{1}{\alpha}\|\mathbf{\Omega}_1 - \mathbf{\Omega}_2\|_2 \tag{6.19}$$

Proof. First, we show that $f(\mathbf{\Omega}_1, \mathbf{\Omega}_2) = \mathcal{KL}(\mathcal{P}_1||\mathcal{P}_2)$ is Lipschitz continuous with respect to $\mathbf{\Omega}_2$. By deriving $\partial f/\partial\mathbf{\Omega}_2 = \frac{1}{2}(-\mathbf{\Omega}_2^{-1} + \mathbf{\Omega}_1^{-1})$. Therefore $\|\partial f/\partial\mathbf{\Omega}_2\|_2 \leq \frac{1}{2}(\|\mathbf{\Omega}_2^{-1}\|_2 + \|\mathbf{\Omega}_1^{-1}\|_2) \leq \frac{1}{2}(\frac{1}{\alpha} + \frac{1}{\alpha}) = \frac{1}{\alpha}$.

Second, since f is Lipschitz continuous with respect to its second parameter, we have $(\forall\mathbf{\Omega})\ |f(\mathbf{\Omega}, \mathbf{\Omega}_2) - f(\mathbf{\Omega}, \mathbf{\Omega}_1)| \leq \frac{1}{\alpha}\|\mathbf{\Omega}_2 - \mathbf{\Omega}_1\|_2$. In particular, let $\mathbf{\Omega} = \mathbf{\Omega}_1$ and since $f(\mathbf{\Omega}_1, \mathbf{\Omega}_1) = 0$ and $|f(\mathbf{\Omega}_1, \mathbf{\Omega}_2)| = f(\mathbf{\Omega}_1, \mathbf{\Omega}_2)$ by properties of the Kullback-Leibler divergence, we prove our claim. □

We also analyze parametrization by using covariance matrices (please, see Appendix J). We point out to the reader that this parametrization does not correspond to factors graphs as in eq.(6.15) and therefore conditional independence does not correspond to zeros in the covariance matrix.

6.4.7 Dynamic Models

The following lemma shows that dynamic Bayesian networks are Lipschitz continuous. Note that dynamic Bayesian networks only impose constraints on the topology of directed graphs, and therefore the extension to the dynamic case is trivial.

Lemma 6.24. *Let $\mathbf{x}^{(t)}$ be the value for variable $\mathbf{x}$ at time t, and let $\mathbf{x}^{(t,\dots,t-L)}$ be a shorthand notation that includes the current time step and the previous L time steps, i.e. $\mathbf{x}^{(t)}, \dots, \mathbf{x}^{(t-L)}$. Let the set of parents for $x_n^{(t)}$ be $\pi_n \subseteq \{1, \dots, N\} \times \{0, \dots, L\}$. Given a (ℓ_p, K)-Lipschitz continuous conditional probability function $p(x_n^{(t)}|\mathbf{x}_{\pi_n}^{(t,\dots,t-L)}, \mathbf{\Theta})$ for each variable $x_n^{(t)}$, the L-order Bayesian network $p(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}, \dots, \mathbf{x}^{(t-L)}, \mathbf{\Theta}) = \prod_n p(x_n^{(t)}|\mathbf{x}_{\pi_n}^{(t,\dots,t-L)}, \mathbf{\Theta})$ is (ℓ_p, NK)-Lipschitz continuous.*

Proof. Similar to proof of Lemma 6.12. □

The following lemma establishes Lipschitz continuity for conditional random fields.

Lemma 6.25. *Given a feature function with F features $\boldsymbol{\psi}(\mathbf{y}, \mathbf{x}) = (\psi_1(\mathbf{y}, \mathbf{x}), ..., \psi_F(\mathbf{y}, \mathbf{x}))^{\mathrm{T}}$, the conditional random field parameterized by $\mathbf{\Theta} = \mathbf{w}$, $\mathbf{w} \in \mathbb{R}^F$ with probability distribution:*

$$p(\mathbf{y}|\mathbf{x}, \mathbf{\Theta}) = \frac{1}{\mathcal{Z}(\mathbf{x}, \mathbf{w})} e^{\mathbf{w}^{\mathrm{T}} \boldsymbol{\psi}(\mathbf{y}, \mathbf{x})} \tag{6.20}$$

where $\mathcal{Z}(\mathbf{x}, \mathbf{w}) = \int_{\mathbf{y}} e^{\mathbf{w}^{\mathrm{T}} \boldsymbol{\psi}(\mathbf{y}, \mathbf{x})}$ is (ℓ_p, K)-Lipschitz continuous.

Proof. Similar to proof of Lemma 6.19 for discrete random variables, or Lemma 6.20 for continuous random variables. □

6.5 Experimental Results

First, we show the similarities between the Kullback-Leibler divergence, test log-likelihood and Frobenius norm for some probabilistic graphical models: Gaussian graphical models for continuous data and Ising models for discrete data. Note that if we assume that the test data is generated by a ground truth model, the expected value of the test log-likelihood is the expected log-likelihood that we analyzed in Section 6.3.

Gaussian graphical models were parameterized by their precision matrices as in eq.(6.17). We consider Ising models without external field. Therefore, in both cases conditional independence corresponds to parameters of value zero. The ground truth model contains $N = 50$ variables for Gaussian graphical models. For Ising models, since computing the log-partition function is NP-hard, we restrict our experiments to $N = 10$ variables. For each of 50 repetitions, we generate edges in the ground truth model with a required density (either 0.2,0.5,0.8), where each edge weight is generated uniformly at random from $[-1; +1]$. For Gaussian graphical models, we ensure positive definiteness by verifying that the minimum eigenvalue is at least 0.1. We then generate training and testing datasets of 50 samples each. Gaussian graphical models were learnt by the graphical lasso method of Friedman et al. [2007b], and Ising models were learnt by the pseudolikelihood method of Höfling and

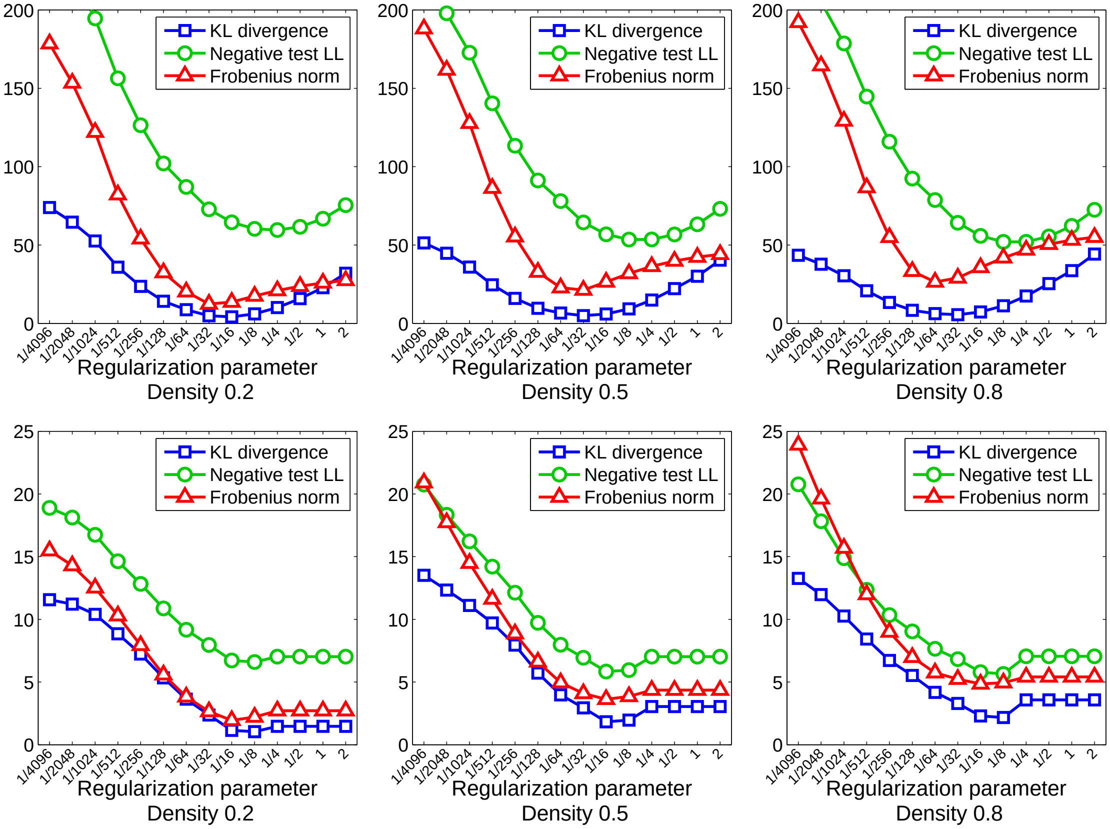


Figure 6.1: Kullback-Leibler divergence, negative test log-likelihood and Frobenius norm for Gaussian graphical models (top) and Ising models (bottom), for low (left) moderate (center) and high (right) graph density. Note that all the measurements behave similarly.

Tibshirani [2009]. Figure 6.1 shows that the Kullback-Leibler divergence, negative test log-likelihood and Frobenius norm behave similarly.

Next, we test the usefulness of our theoretical results that enable us to perform classification, dimensionality reduction and clustering from the parameters of graphical models. We use the CMU motion capture database (`http://mocap.cs.cmu.edu/`) for activity recognition and temporal segmentation. In both cases, we only used the Euler angles for the following 8 markers: left and right humerus, radius, femur and tibia. Our variables measure the change in Euler angles, i.e. the difference between the angle at the current time and 0.05 seconds before. Variables were normalized to have standard deviation one.

For activity recognition, we test whether it is possible to detect if a person is either walking or running from a small window of 0.25 seconds (through the use of classification). The CMU motion capture database contains several sequences per subject. We used the first sequence labeled as "walk" or "run" from all available subjects (excluding 3 pregnant and post-pregnant women). This led to 14 walking subjects and 10 running subjects (total of 21 distinct subjects). From each subject we extracted 3 small windows of 0.25 seconds, at 1/4, 2/4 and 3/4 of the whole sequence. Covariance and precision matrices of Gaussian graphical models were learnt by Tikhonov regularization and the covariance selection method of Banerjee et al. [2006]. Table 6.1 shows the leave-one-subject-out accuracy for a linear SVM

Table 6.1: Leave-one-subject-out accuracy for walking vs. running on the CMU motion capture database (chance = 58%).

Regularization level	0.001	0.01	0.1	1
ℓ_1 Covariance	78	78	74	76
Tikhonov Covariance	78	78	78	78
ℓ_1 Precision	96	93	90	75
Tikhonov Precision	97	96	93	92

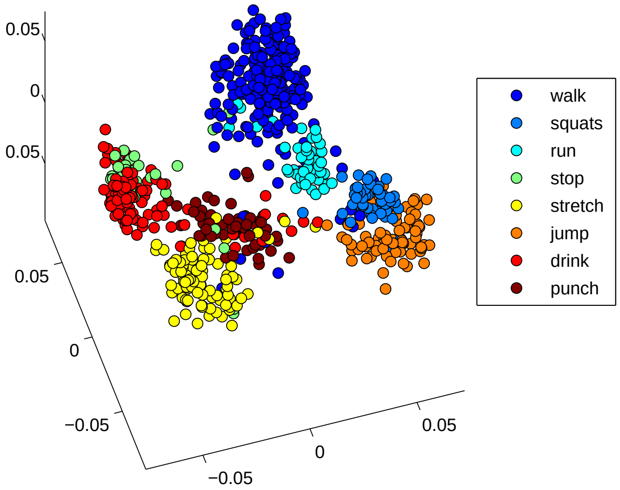


Figure 6.2: Clusters from a complex sequence of the CMU motion capture database. Each point represents a Gaussian graphical model, the Kullback-Leibler divergence between two points is bounded by the distance between them.

classifier with the parameters of the Gaussian graphical models as features.

For temporal segmentation, we test whether it is possible to separate a complex sequence that includes walking, squats, running, stopping, stretching, jumping, drinking and punching (through dimensionality reduction and clustering). We used the sequence 2 of subject 86 from the CMU motion capture database. We extracted small windows of 0.75 seconds, taken each 0.125 seconds. Each window was labeled as the action being executed in the middle. Precision matrices of Gaussian graphical models were learnt by Tikhonov regularization with regularization level 0.1. We first apply PCA by using the parameters of the Gaussian graphical models as features and then perform k-means clustering with the first 3 eigenvectors. Figure 6.2 shows the resulting clusters and Table 6.2 shows the confusion matrix of assigning each window to its cluster.

6.6 Concluding Remarks

One of our contributions was to show that methods that penalize the ℓ_p-norm of differences of parameters [Kolar et al., 2009, 2010, Zhang and Wang, 2010] are minimizing an upper bound of the Kullback-Leibler divergence. Along this line, we can further discuss the role of the ℓ_1-regularization (e.g. [Banerjee et al., 2006, Friedman et al., 2007b, Lee et al., 2006a, Schmidt

Table 6.2: Confusion matrix for temporal segmentation from a complex sequence of the CMU motion capture database. Ground truth labels on each row, predicted labels on each column (each row add up to 100%).

	walk	squats	run	stop	stretch	jump	drink	punch
walk	93				1	2		4
squats		87				7		6
run	13		83					4
stop	6			73	3		3	15
stretch					70			30
jump		4				96		
drink				6	4	1	78	11
punch					16		4	80

et al., 2007b, Wainwright et al., 2006]), which has been used for promoting sparseness and that it is equivalent to a Laplacian prior. We can argue that ℓ_1-regularization imposes a prior that reduces the Kullback-Leibler divergence between the learnt model and the independent model. For instance, in the case of factor graphs, assume a learnt model with weights $\mathbf{b}$ for unitary potentials, and weights $\mathbf{W}$ for non-unitary potentials (e.g. pairs, triplets), then the Kullback-Leibler divergence between the learnt model $(\mathbf{W}, \mathbf{b})$ and the independent model $(\mathbf{0}, \mathbf{b})$ is bounded by a term that is $\mathcal{O}(\|(\mathbf{W}, \mathbf{b}) - (\mathbf{0}, \mathbf{b})\|_1) = \mathcal{O}(\|\mathbf{W}\|_1)$.

There are several ways of extending this research. Lipschitz continuity for the parameterization of other probability distributions (e.g. mixture models) needs to be analyzed. We hope that our preliminary results will motivate work on proving other theoretical properties as well as on learning probabilistic graphical models by using optimization algorithms that rely on Lipschitz continuity of the log-likelihood as the objective function. Finally, while Lipschitz continuity defines an upper bound of the derivative, lower bounds of the derivative will allow for finding a lower bound of the Kullback-Leibler divergence as well as upper bounds for the Bayes error and the expected log-likelihood.

Chapter 7

Learning Linear Influence Games

In the previous chapters, we focused on probabilistic graphical models, in which graphs encode conditional dependence relationships. In this chapter, we focus on graphical games, in which graphs encode *strategic* dependence relationships.

We formalize and study the problem of learning the structure and parameters of *graphical games* from strictly *behavioral* data. We cast the problem as a maximum likelihood estimation based on a generative model defined by the *pure-strategy Nash equilibria* of the game. The formulation brings out the interplay between goodness-of-fit and model complexity: good models capture the equilibrium behavior represented in the data while controlling the *true* number of equilibria, including those potentially unobserved. We provide a generalization bound for maximum likelihood estimation. We discuss several optimization algorithms including *convex loss minimization*, sigmoidal approximations and exhaustive search. We formally prove that games in our hypothesis space have a small *true* number of equilibria, with high probability; thus, convex loss minimization is sound. We illustrate our approach, show and discuss promising results on synthetic data and the U.S. congressional voting records.

7.1 Introduction

Graphical games [Kearns et al., 2001] were one of the first and most influential graphical models for game theory. It has been about a decade since their introduction to the AI community. There has also been considerable progress on problems of *computing* classical equilibrium solution concepts such as Nash [Nash, 1951] and correlated equilibria [Aumann, 1974] in graphical games (see, e.g., Kearns et al. [2001], Vickrey and Koller [2002], Ortiz and Kearns [2002], Blum et al. [2006], Kakade et al. [2003], Papadimitriou and Roughgarden [2008], Jiang and Leyton-Brown [2011] and the references therein). Indeed, graphical games played a prominent role in establishing the computational complexity of computing Nash equilibria in general normal-form games (see, e.g., Daskalakis et al. [2009] and the references therein).

Relatively less attention has been paid to the problem of *learning* the structure of graphical games from data. Addressing this problem is essential to the development, potential use and success of game-theoretic models in practical applications.

Indeed, we are beginning to see an increase in the availability of data collected from processes that are the result of deliberate actions of agents in complex system. A lot of this data results from the interaction of a large number of individuals, being people, companies, governments, groups or engineered autonomous systems (e.g. autonomous trading agents), for which any form of global control is usually weak. The Internet is currently a major source of such data, and the smart grid, with its trumpeted ability to allow individual customers to install autonomous control devices and systems for electricity demand, will likely be another one in the near future.

We present a formal framework and design algorithms for learning the structure and parameters of graphical games [Kearns et al., 2001] in large populations of agents. We concentrate on learning from purely behavioral data. We expect that, in most cases, the parameters quantifying a utility function or best-response condition are unavailable and hard to determine in real-world settings. The availability of data resulting from the observation of an individual *public behavior* is arguably a weaker assumption than the availability of individual *utility* observations, which are often *private*.

Our technical contributions include a novel generative model of behavioral data in Section 7.4 for general games. We define identifiability and triviality of games. We provide conditions which ensures identifiability among non-trivial games. We then present the maximum likelihood problem for general (non-trivial identifiable) games. In Section 7.5, we show a generalization bound for the maximum likelihood problem as well as an upper bound of the VC-dimension of influence games. In Section 7.6, we approximate the original problem by maximizing the number of observed equilibria in the data, suitable for a hypothesis space of games with small *true* number of equilibria. We then present our convex loss minimization approach and a baseline sigmoidal approximation for (linear) influence games. We also present exhaustive search methods for both general as well as influence games. In Section 7.7, we define absolute-indifference of players and show that our convex loss minimization approach produces games in which all players are non-absolutely-indifferent. We provide a distribution-free bound which shows that linear influence games have small *true* number of equilibria with high probability.

7.2 Related Work

Our work *complements* the recent line of work on learning graphical games [Vorobeychik et al., 2005, Ficici et al., 2008, Duong et al., 2009, Gao and Pfeffer, 2010, Ziebart et al., 2010, Waugh et al., 2011]. With the exception of Ziebart et al. [2010], Waugh et al. [2011], previous methods assume that the actions as well as corresponding payoffs (or noisy samples from the true payoff function) are observed in the data. Another notable exception is a recently proposed framework from the learning theory community to model *collective* behavior [Kearns and Wortman, 2008]. The approach taken there considers dynamics and is based on stochastic models. Our work differs from methods that assume that the game is known [Wright and Leyton-Brown, 2010]. The work of Vorobeychik et al. [2005], Gao and Pfeffer [2010], Wright and Leyton-Brown [2010], Ziebart et al. [2010] present experimental validation mostly for 2 players only, 7 players in Waugh et al. [2011] and up to 13 players in Duong et al. [2009].

In this chapter, we assume that the joint-actions is the only observable information. To the best of our knowledge, we present the first techniques for learning the structure and parameters of large-population graphical games from joint-actions only. Furthermore, we present experimental validation in games of up to 100 players. Our convex loss minimization approach could potentially be applied to larger problems since it is polynomial-time.

There has been a significant amount of work for learning the structure of *probabilistic* graphical models from data. We mention only a few references that follow a maximum likelihood approach for Markov random fields [Lee et al., 2006a], bounded tree-width distributions [Chow and Liu, 1968, Srebro, 2001], Ising models [Wainwright et al., 2006, Banerjee et al., 2008, Höfling and Tibshirani, 2009], Gaussian graphical models [Banerjee et al., 2006], Bayesian networks [Guo and Schuurmans, 2006, Schmidt et al., 2007b] and directed cyclic graphs [Schmidt and Murphy, 2009].

Our approach learns the structure and parameters of games by maximum likelihood estimation on a related probabilistic model. Our probabilistic model does not fit into any of the types described above. Although a (directed) graphical game has a directed cyclic graph, there is a semantic difference with respect to graphical models. Structure in a graphical model implies a factorization of the probabilistic model. In a graphical game, the graph structure implies *strategic* dependence between players, and has no immediate probabilistic implication. Furthermore, our general model differs from Schmidt and Murphy [2009] since our generative model does not decompose as a multiplication of potential functions.

7.3 Background

In classical game-theory (see, e.g. Fudenberg and Tirole [1991] for a textbook introduction), a *normal-form game* is defined by a set of *players* V (e.g. we can let $V = \{1, \ldots, n\}$ if there are n players), and for each player i, a set of *actions*, or *pure-strategies* A_i, and a payoff function $u_i : \times_{j \in V} A_j \to \mathbb{R}$ mapping the joint-actions of all the players, given by the Cartesian product $\mathcal{A} \equiv \times_{j \in V} A_j$, to a real number. In non-cooperative game theory we assume players are greedy, rational and act independently, by which we mean that each player i always want to maximize their own utility, subject to the actions selected by others, irrespective of how the optimal action chosen help or hurt others.

A core solution concept in non-cooperative game theory is that of an *Nash equilibrium.* A joint-action $\mathbf{x}^* \in \mathcal{A}$ is a *pure-strategy Nash equilibrium* of a non-cooperative game if, for each player i, $x_i^* \in \arg\max_{x_i \in A_i} u_i(x_i, \mathbf{x}_{-i}^*)$; that is, $\mathbf{x}^*$ constitutes a *mutual best-response*, no player i has any incentive to unilaterally deviate from the prescribed action x_i^*, given the joint-action of the other players $\mathbf{x}_{-i}^* \in \times_{j \in V - \{i\}} A_j$ in the equilibrium.

In what follows, we denote a game by $\mathcal{G}$, and the set of all *pure-strategy Nash equilibria* of $\mathcal{G}$ by:

$$\mathcal{NE}(\mathcal{G}) \equiv \{\mathbf{x}^* \mid (\forall i \in V)\ x_i^* \in \arg\max_{x_i \in A_i} u_i(x_i, \mathbf{x}_{-i}^*)\} \tag{7.1}$$

A *(directed) graphical game* is a game-theoretic graphical model [Kearns et al., 2001]. It provides a succinct representation of normal-form games. In a graphical game, we have a (directed) graph $G = (V, E)$ in which each node in V corresponds to a player in the game. The interpretation of the edges/arcs E of G is that the payoff function of player i is only a function of the set of parents/neighbors $\mathcal{N}_i \equiv \{j \mid (i, j) \in E\}$ in G (i.e. the set of players corresponding to nodes that point to the node corresponding to player i in the graph). In the context of a graphical game, we refer to the u_i's as the *local payoff functions/matrices.*

Linear influence games [Irfan and Ortiz, 2011] are a sub-class of graphical games. For linear influence games, we assume that we are given a matrix of influence weights $\mathbf{W} \in \mathbb{R}^{n \times n}$, with zero diagonal (i.e. $\mathbf{diag}(\mathbf{W}) = \mathbf{0}$), and a threshold vector $\mathbf{b} \in \mathbb{R}^n$. For each player i, we define the influence function $f_i(\mathbf{x}_{-i}) \equiv \sum_{j \in \mathcal{N}_i} w_{ij} x_j - b_i = \mathbf{w}_{i,-i}{}^{\mathrm{T}} \mathbf{x}_{-i} - b_i$ and the payoff function $u_i(\mathbf{x}) \equiv x_i f_i(\mathbf{x}_{-i})$. We further assume binary actions: $A_i \equiv \{-1, +1\}$ for all i. The *best response* x_i^* of player i to the joint-action $\mathbf{x}_{-i}$ of the other players is defined as:

$$\left\{ \begin{array}{l} \mathbf{w}_{i,-i}{}^{\mathrm{T}} \mathbf{x}_{-i} > b_i \Rightarrow x_i^* = +1, \\ \mathbf{w}_{i,-i}{}^{\mathrm{T}} \mathbf{x}_{-i} < b_i \Rightarrow x_i^* = -1 \text{ and} \\ \mathbf{w}_{i,-i}{}^{\mathrm{T}} \mathbf{x}_{-i} = b_i \Rightarrow x_i^* \in \{-1, +1\} \end{array} \right\} \Leftrightarrow x_i^* (\mathbf{w}_{i,-i}{}^{\mathrm{T}} \mathbf{x}_{-i} - b_i) \geq 0 \tag{7.2}$$

Hence, for any other player j, $w_{ij} \in \mathbb{R}$ can be thought as a *weight* parameter quantifying the "influence factor" that j has on i, and $b_i \in \mathbb{R}$ as a *threshold* parameter to the level of "tolerance" that player i has for playing -1.

As discussed in Irfan and Ortiz [2011], linear influence games are also a sub-class of polymatrix games [Janovskaja, 1968]. Furthermore, in the special case of $\mathbf{b} = \mathbf{0}$ and symmetric $\mathbf{W}$, a linear influence game becomes a *party-affiliation game* [Fabrikant et al., 2004].

Figure 7.3 provides a preview illustration of the application of our approach to congressional voting.

7.4 Preliminaries

Our goal is to learn the structure and parameters of a graphical game from observed joint-actions. Note that our problem is unsupervised, i.e. we do not know a priori which joint-actions are equilibria and which ones are not. If our only goal were to find a game $\mathcal{G}$ in which all the given observed data is an equilibrium, then a "dummy" influence game with $\mathcal{G} = (\mathbf{W}, \mathbf{b}), \mathbf{W} = \mathbf{0}, \mathbf{b} = \mathbf{0}$ would be the optimal solution since $|\mathcal{NE}(\mathcal{G})| = 2^n$. In this section, we present a probabilistic formulation that allows finding games that maximize the *empirical proportion of equilibria* in the data while keeping the *true proportion of equilibria* as low as possible. Furthermore, we show that *trivial* games such as $\mathbf{W} = \mathbf{0}, \mathbf{b} = \mathbf{0}$, obtain the lowest log-likelihood.

7.4.1 On the Identifiability of Games

Several games with different coefficients can lead to the same Nash equilibria set. As a simple example that illustrates the issue of identifiability, consider the three following influence games with the same Nash equilibria sets, i.e. $\mathcal{NE}(\mathbf{W}_k, \mathbf{0}) = \{(-1, -1, -1), (+1, +1, +1)\}$ for $k = 1, 2, 3$:

$$\mathbf{W}_1 = \begin{bmatrix} 0 & 0 & 0 \\ 1/2 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad \mathbf{W}_2 = \begin{bmatrix} 0 & 0 & 0 \\ 2 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \quad \mathbf{W}_3 = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

Clearly, using structural properties alone, one would generally prefer the former two models to the latter, all else being equal (e.g. generalization performance). A large number of the

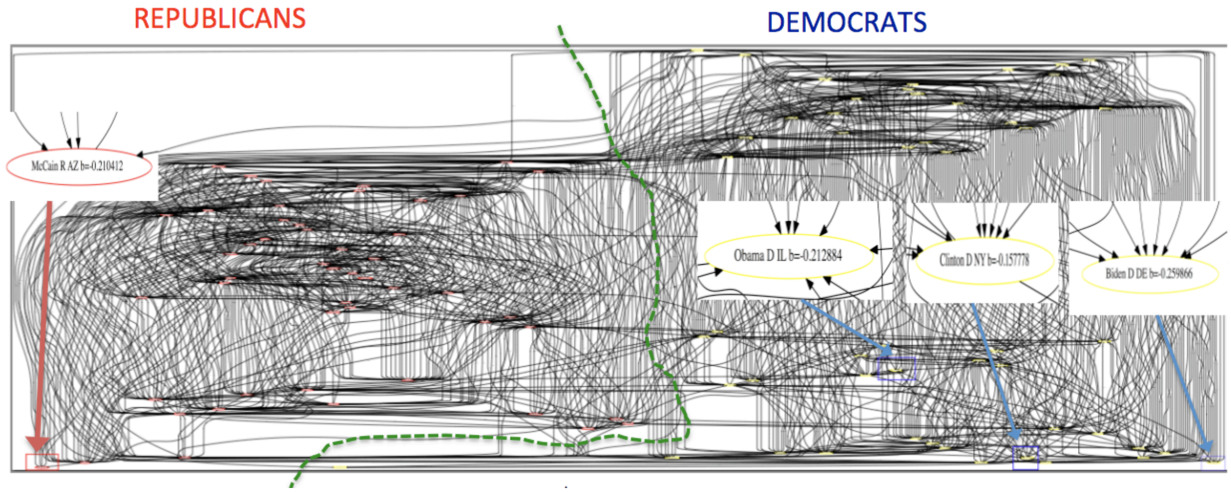


Figure 7.1: **110th US Congress's Linear Influence Game (January 3, 2007-09):** We provide an illustration of the application of our approach to real congressional voting data. Irfan and Ortiz [2011] use such LIGs to address a variety of computational problems, including the identification of *most influential* senators. We show the graph connectivity of a LIG learnt by independent ℓ_1-regularized logistic regression (see Sect. 7.6.5). We highlight some characteristics of the graph, consistent with anecdotal evidence. First, senators are more likely to be influenced by members of the same party than by members of the opposite party (the dashed green line denotes the separation between the parties). Republicans were "more strongly united" (tighter connectivity) than Democrats at the time. Second, the current US Vice President Biden (Dem./Delaware) and McCain (Rep./Arizona) are displayed at the "extreme of each party" (Biden at the bottom-right corner, McCain at the bottom-left) eliciting their opposite ideologies. Third, note that Biden, McCain, the current US President Obama (Dem./Illinois) and US Secretary of State Hillary Clinton (Dem./New York) have very few outgoing arcs; e.g., Obama only directly influences Feingold (Dem./Wisconsin), a prominent senior member with strongly liberal stands. One may wonder why do such prominent senators seem to have so little direct influence on others? A possible explanation is that US President Bush was about to complete hist second term (the maximum allowed). Both parties had *very long* presidential primaries. All those senators contended for the presidential candidacy within their parties. Hence, one may posit that those senators were focusing on running their campaigns and that their influence in the *day-to-day* business of congress was channeled through other prominent senior members of their parties.

econometrics literature concerns the issue of identifiability of models from data. In typical machine-learning fashion, we side-step this issue by measuring the quality of our data-induced models via their generalization ability and invoke the principle of Ockham's razor to bias our search toward simpler models using well-known and -studied regularization techniques. In particular, we take the view that games are identifiable by their Nash equilibria. Hence our next definition.

Definition 7.1. *We say that two games $\mathcal{G}_1$ and $\mathcal{G}_2$ are* equivalent *if and only if their Nash equilibria sets are identical, i.e.: $\mathcal{G}_1 \equiv_{\mathcal{NE}} \mathcal{G}_2 \Leftrightarrow \mathcal{NE}(\mathcal{G}_1) = \mathcal{NE}(\mathcal{G}_2)$.*

7.4.2 Generative Model of Behavioral Data

We propose the following generative model for behavioral data based strictly in the context of "simultaneous"/one-shot play in non-cooperative game theory. Let $\mathcal{G}$ be a game. With some probability $0 < q < 1$, a joint-action $\mathbf{x}$ is chosen uniformly at random from $\mathcal{NE}(\mathcal{G})$; otherwise, $\mathbf{x}$ is chosen uniformly at random from its complement set $\{-1,+1\}^n - \mathcal{NE}(\mathcal{G})$. Hence, the generative model is a mixture model with mixture parameter q corresponding to the probability that a stable outcome (i.e. a Nash equilibrium) of the game is observed. Formally, the probability mass function (PMF) over joint-behaviors $\{-1,+1\}^n$ parametrized by $(\mathcal{G}, q)$ is:

$$p_{(\mathcal{G},q)}(\mathbf{x}) = q\,\frac{1[\mathbf{x} \in \mathcal{NE}(\mathcal{G})]}{|\mathcal{NE}(\mathcal{G})|} + (1-q)\,\frac{1[\mathbf{x} \notin \mathcal{NE}(\mathcal{G})]}{2^n - |\mathcal{NE}(\mathcal{G})|} \tag{7.3}$$

where we can think of q as the "signal" level, and thus $1-q$ as the "noise" level in the data set.

Remark 7.2. *Note that in order for eq.*(7.3) *to be a valid PMF for any* $\mathcal{G}$*, we need to enforce the following conditions* $|\mathcal{NE}(\mathcal{G})| = 0 \Rightarrow q = 0$ *and* $|\mathcal{NE}(\mathcal{G})| = 2^n \Rightarrow q = 1$*. Furthermore, note that in both cases* ($|\mathcal{NE}(\mathcal{G})| \in \{0, 2^n\}$) *the PMF becomes a uniform distribution. On the other hand, if* $0 < |\mathcal{NE}(\mathcal{G})| < 2^n$ *then setting* $q \in \{0,1\}$ *leads to an invalid PMF.*

Let $\pi(\mathcal{G})$ be the *true proportion of equilibria* in the game $\mathcal{G}$ relative to all possible joint-actions, i.e.:

$$\pi(\mathcal{G}) \equiv |\mathcal{NE}(\mathcal{G})|/2^n \tag{7.4}$$

Definition 7.3. *We say that a game* $\mathcal{G}$ *is trivial if and only if* $|\mathcal{NE}(\mathcal{G})| \in \{0, 2^n\}$ *(or equivalently* $\pi(\mathcal{G}) \in \{0,1\}$*), and non-trivial if and only if* $0 < |\mathcal{NE}(\mathcal{G})| < 2^n$ *(or equivalently* $0 < \pi(\mathcal{G}) < 1$*).*

The following propositions establish that the condition $q > \pi(\mathcal{G})$ ensures that the probability of an equilibrium is strictly greater than a non-equilibrium. The condition also guarantees identifiability among non-trivial games.

Proposition 7.4. *Given a non-trivial game* $\mathcal{G}$*, the mixture parameter* $q > \pi(\mathcal{G})$ *if and only if* $p_{(\mathcal{G},q)}(\mathbf{x}_1) > p_{(\mathcal{G},q)}(\mathbf{x}_2)$ *for any* $\mathbf{x}_1 \in \mathcal{NE}(\mathcal{G})$ *and* $\mathbf{x}_2 \notin \mathcal{NE}(\mathcal{G})$*.*

Proof. Note that $p_{(\mathcal{G},q)}(\mathbf{x}_1) = q/|\mathcal{NE}(\mathcal{G})| > p_{(\mathcal{G},q)}(\mathbf{x}_2) = (1-q)/(2^n - |\mathcal{NE}(\mathcal{G})|) \Leftrightarrow q > |\mathcal{NE}(\mathcal{G})|/2^n$ and given eq.(7.4), we prove our claim. □

Proposition 7.5. *Let* $\mathcal{G}_1$ *and* $\mathcal{G}_2$ *be two non-trivial games. For some mixture parameter* $q > \max(\pi(\mathcal{G}_1), \pi(\mathcal{G}_2))$*,* $\mathcal{G}_1$ *and* $\mathcal{G}_2$ *are* equivalent *if and only if they induce the same PMF over the joint-action space* $\{-1,+1\}^n$ *of the players, i.e.:* $\mathcal{G}_1 \equiv_{\mathcal{NE}} \mathcal{G}_2 \Leftrightarrow (\forall \mathbf{x})\ p_{(\mathcal{G}_1,q)}(\mathbf{x}) = p_{(\mathcal{G}_2,q)}(\mathbf{x})$*.*

Proof. Let $\mathcal{NE}_k \equiv \mathcal{NE}(\mathcal{G}_k)$. First, we prove the $\Rightarrow$ direction. By Definition 7.1, $\mathcal{G}_1 \equiv_{\mathcal{NE}} \mathcal{G}_2 \Rightarrow \mathcal{NE}_1 = \mathcal{NE}_2$. Note that $p_{(\mathcal{G}_k,q)}(\mathbf{x})$ in eq.(7.3) depends only on characteristic functions $1[\mathbf{x} \in \mathcal{NE}_k]$. Therefore, $(\forall \mathbf{x})\ p_{(\mathcal{G}_1,q)}(\mathbf{x}) = p_{(\mathcal{G}_2,q)}(\mathbf{x})$.

Second, we prove the $\Leftarrow$ direction by contradiction. Assume $(\exists \mathbf{x})\ \mathbf{x} \in \mathcal{NE}_1 \wedge \mathbf{x} \notin \mathcal{NE}_2$. $p_{(\mathcal{G}_1,q)}(\mathbf{x}) = p_{(\mathcal{G}_2,q)}(\mathbf{x})$ implies that $q/|\mathcal{NE}_1| = (1-q)/(2^n - |\mathcal{NE}_2|) \Rightarrow q = |\mathcal{NE}_1|/(2^n + |\mathcal{NE}_1| -$

$|\mathcal{NE}_2|)$. Since $q > \max(\pi(\mathcal{G}_1), \pi(\mathcal{G}_2)) \Rightarrow q > \max(|\mathcal{NE}_1|, |\mathcal{NE}_2|)/2^n$ by eq.(7.4). Therefore $\max(|\mathcal{NE}_1|, |\mathcal{NE}_2|)/2^n < |\mathcal{NE}_1|/(2^n + |\mathcal{NE}_1| - |\mathcal{NE}_2|)$. If we assume that $|\mathcal{NE}_1| \geq |\mathcal{NE}_2|$ we reach the contradiction $|\mathcal{NE}_1| - |\mathcal{NE}_2| < 0$. If we assume that $|\mathcal{NE}_1| \leq |\mathcal{NE}_2|$ we reach the contradiction $(2^n - |\mathcal{NE}_2|)(|\mathcal{NE}_2| - |\mathcal{NE}_1|) < 0$. □

Remark 7.6. *Recall that a trivial game induces a uniform PMF by Remark 7.2. Therefore, a non-trivial game is not* equivalent *to a trivial game since by Proposition 7.4, non-trivial games do not induce uniform PMFs.*

7.4.3 Learning the Structure of Games via Maximum Likelihood Estimation

The *learning problem* consists on estimating the structure and parameters of a graphical game from data. We point out that our problem is unsupervised, i.e. we do not know a priori which joint-actions are equilibria and which ones are not. We based our framework on the fact that games are identifiable with respect to their induced PMF by Proposition 7.5.

First, we introduce a shorthand notation for the Kullback-Leibler (KL) divergence between two Bernoulli distributions parametrized by $0 \leq p_1 \leq 1$ and $0 \leq p_2 \leq 1$:

$$\begin{aligned} KL(p_1 \| p_2) &\equiv KL(\text{Bernoulli}(p_1) \| \text{Bernoulli}(p_2)) \\ &= p_1 \log \tfrac{p_1}{p_2} + (1 - p_1) \log \tfrac{1-p_1}{1-p_2} \end{aligned} \tag{7.5}$$

Using this function, we can derive the following expression of the maximum likelihood estimation problem.

Lemma 7.7. *Given a dataset $\mathcal{D} = \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$, let $\widehat{\pi}(\mathcal{G})$ be the* empirical proportion of equilibria, *i.e. the proportion of samples in $\mathcal{D}$ that are equilibria of $\mathcal{G}$:*

$$\widehat{\pi}(\mathcal{G}) \equiv \tfrac{1}{m} \textstyle\sum_l 1[\mathbf{x}^{(l)} \in \mathcal{NE}(\mathcal{G})] \tag{7.6}$$

the maximum likelihood estimation problem for the probabilistic model in eq.(7.3) can be expressed as:

$$\max_{(\mathcal{G},q) \in \Upsilon} \widehat{\mathcal{L}}(\mathcal{G}, q) \quad , \quad \widehat{\mathcal{L}}(\mathcal{G}, q) = KL(\widehat{\pi}(\mathcal{G}) \| \pi(\mathcal{G})) - KL(\widehat{\pi}(\mathcal{G}) \| q) - n \log 2 \tag{7.7}$$

where $\mathcal{H}$ is the class of games of interest, $\Upsilon = \{(\mathcal{G}, q) \mid \mathcal{G} \in \mathcal{H} \wedge 0 < \pi(\mathcal{G}) < q < 1\}$ is the hypothesis space of non-trivial identifiable games, $\pi(\mathcal{G})$ is defined as in eq.(7.4) and the optimal mixture parameter $\widehat{q} = \min(\widehat{\pi}(\mathcal{G}), 1 - \frac{1}{2m})$.

Proof. Let $\mathcal{NE} \equiv \mathcal{NE}(\mathcal{G})$, $\pi \equiv \pi(\mathcal{G})$ and $\widehat{\pi} \equiv \widehat{\pi}(\mathcal{G})$. First, for a non-trivial $\mathcal{G}$, $\log p_{(\mathcal{G},q)}(\mathbf{x}^{(l)}) = \log \frac{q}{|\mathcal{NE}|}$ for $\mathbf{x}^{(l)} \in \mathcal{NE}$, and $\log p_{(\mathcal{G},q)}(\mathbf{x}^{(l)}) = \log \frac{1-q}{2^n - |\mathcal{NE}|}$ for $\mathbf{x}^{(l)} \notin \mathcal{NE}$. The average log-likelihood $\widehat{\mathcal{L}}(\mathcal{G}, q) = \frac{1}{m} \sum_l \log p_{\mathcal{G},q}(\mathbf{x}^{(l)}) = \widehat{\pi} \log \frac{q}{|\mathcal{NE}|} + (1 - \widehat{\pi}) \log \frac{1-q}{2^n - |\mathcal{NE}|} = \widehat{\pi} \log \frac{q}{\pi} + (1 - \widehat{\pi}) \log \frac{1-q}{1-\pi} - n \log 2$. By adding $0 = -\widehat{\pi} \log \widehat{\pi} + \widehat{\pi} \log \widehat{\pi} - (1 - \widehat{\pi}) \log(1 - \widehat{\pi}) + (1 - \widehat{\pi}) \log(1 - \widehat{\pi})$, this can be rewritten as $\widehat{\mathcal{L}}(\mathcal{G}, q) = \widehat{\pi} \log \frac{\widehat{\pi}}{\pi} + (1 - \widehat{\pi}) \log \frac{1-\widehat{\pi}}{1-\pi} - \widehat{\pi} \log \frac{\widehat{\pi}}{q} - (1 - \widehat{\pi}) \log \frac{1-\widehat{\pi}}{1-q} - n \log 2$, and by using eq.(7.5) we prove our claim.

Note that by maximizing with respect to the mixture parameter q and by properties of the KL divergence, we get $KL(\widehat{\pi}\|\widehat{q}) = 0 \Leftrightarrow \widehat{q} = \widehat{\pi}$. We define our hypothesis space Υ given the conditions in Remark 7.2 and Propositions 7.4 and 7.5. For the case $\widehat{\pi} = 1$, we "shrink" the optimal mixture parameter $\widehat{q}$ to $1 - \frac{1}{2m}$ in order to avoid generating an invalid PMF as discussed in Remark 7.2. □

Remark 7.8. *Recall that a trivial game (e.g.* $\mathcal{G} = (\mathbf{W}, \mathbf{b}), \mathbf{W} = \mathbf{0}, \mathbf{b} = \mathbf{0}, \pi(\mathcal{G}) = 1$*) induces a uniform PMF by Remark 7.2, and therefore its log-likelihood is* $-n \log 2$*. Note that the lowest log-likelihood for non-trivial identifiable games in eq.(7.7) is* $-n \log 2$ *by setting the optimal mixture parameter* $\widehat{q} = \widehat{\pi}(\mathcal{G})$ *and given that* $KL(\widehat{\pi}(\mathcal{G})\|\pi(\mathcal{G})) \geq 0$.

Furthermore, eq.(7.7) implies that for non-trivial identifiable games $\mathcal{G}$, we expect the *true proportion of equilibria* $\pi(\mathcal{G})$ to be strictly less than the *empirical proportion of equilibria* $\widehat{\pi}(\mathcal{G})$ in the given data. This is by setting the optimal mixture parameter $\widehat{q} = \widehat{\pi}(\mathcal{G})$ and the condition $q > \pi(\mathcal{G})$ in our hypothesis space.

7.5 Generalization Bound and VC-Dimension

In this section, we show a generalization bound for the maximum likelihood problem as well as an upper bound of the VC-dimension of linear influence games. Our objective is to establish that with probability at least $1 - \delta$, for some confidence parameter $0 < \delta < 1$, the maximum likelihood estimate is within $\epsilon > 0$ of the optimal parameters, in terms of achievable expected log-likelihood.

Given the ground truth distribution $\mathcal{Q}$ of the data, let $\bar{\pi}(\mathcal{G})$ be the *expected proportion of equilibria*, i.e.:

$$\bar{\pi}(\mathcal{G}) = \mathbb{P}_{\mathcal{Q}}[\mathbf{x} \in \mathcal{NE}(\mathcal{G})] \tag{7.8}$$

and let $\bar{\mathcal{L}}(\mathcal{G}, q)$ be the *expected log-likelihood* of a generative model from game $\mathcal{G}$ and mixture parameter q, i.e.:

$$\bar{\mathcal{L}}(\mathcal{G}, q) = \mathbb{E}_{\mathcal{Q}}[\log p_{(\mathcal{G},q)}(\mathbf{x})] \tag{7.9}$$

Note that our hypothesis space Υ in eq.(7.7) includes a continuous parameter q that could potentially have infinite VC-dimension. The following lemma will allow us later to prove that uniform convergence for the extreme values of q implies uniform convergence for all q in the domain.

Lemma 7.9. *Consider any game* $\mathcal{G}$ *and, for* $0 < q'' < q' < q < 1$*, let* $\theta = (\mathcal{G}, q)$*,* $\theta' = (\mathcal{G}, q')$ *and* $\theta'' = (\mathcal{G}, q'')$*. If, for any* $\epsilon > 0$ *we have* $|\widehat{\mathcal{L}}(\theta) - \bar{\mathcal{L}}(\theta)| \leq \epsilon/2$ *and* $|\widehat{\mathcal{L}}(\theta'') - \bar{\mathcal{L}}(\theta'')| \leq \epsilon/2$*, then* $|\widehat{\mathcal{L}}(\theta') - \bar{\mathcal{L}}(\theta')| \leq \epsilon/2$.

Proof. Let $\mathcal{NE} \equiv \mathcal{NE}(\mathcal{G})$, $\pi \equiv \pi(\mathcal{G})$, $\widehat{\pi} \equiv \widehat{\pi}(\mathcal{G})$, $\bar{\pi} \equiv \bar{\pi}(\mathcal{G})$, and $\mathbb{E}[\cdot]$ and $\mathbb{P}[\cdot]$ be the expectation and probability with respect to the ground truth distribution $\mathcal{Q}$ of the data.

First note that for any $\theta = (\mathcal{G}, q)$, we have $\bar{\mathcal{L}}(\theta) = \mathbb{E}[\log p_{(\mathcal{G},q)}(\mathbf{x})] = \mathbb{E}[1[\mathbf{x} \in \mathcal{NE}] \log \frac{q}{|\mathcal{NE}|} + 1[\mathbf{x} \notin \mathcal{NE}] \log \frac{1-q}{2^n - |\mathcal{NE}|}] = \mathbb{P}[\mathbf{x} \in \mathcal{NE}] \log \frac{q}{|\mathcal{NE}|} + \mathbb{P}[\mathbf{x} \notin \mathcal{NE}] \log \frac{1-q}{2^n - |\mathcal{NE}|} = \bar{\pi} \log \frac{q}{|\mathcal{NE}|} + (1 - \bar{\pi}) \log \frac{1-q}{2^n - |\mathcal{NE}|} = \bar{\pi} \log \left(\frac{q}{1-q} \cdot \frac{2^n - |\mathcal{NE}|}{|\mathcal{NE}|}\right) + \log \frac{1-q}{2^n - |\mathcal{NE}|} = \bar{\pi} \log \left(\frac{q}{1-q} \cdot \frac{1-\pi}{\pi}\right) + \log \frac{1-q}{1-\pi} - n \log 2$.

Similarly, for any $\theta = (\mathcal{G}, q)$, we have $\widehat{\mathcal{L}}(\theta) = \widehat{\pi}\log\left(\frac{q}{1-q}\cdot\frac{1-\pi}{\pi}\right) + \log\frac{1-q}{1-\pi} - n\log 2$. So that $\widehat{\mathcal{L}}(\theta) - \bar{\mathcal{L}}(\theta) = (\widehat{\pi} - \bar{\pi})\log\left(\frac{q}{1-q}\cdot\frac{1-\pi}{\pi}\right)$.

Furthermore, the function $\frac{q}{1-q}$ is strictly monotonically increasing for $0 \le q < 1$. If $\widehat{\pi} > \bar{\pi}$ then $-\epsilon/2 \le \widehat{\mathcal{L}}(\theta'') - \bar{\mathcal{L}}(\theta'') < \widehat{\mathcal{L}}(\theta') - \bar{\mathcal{L}}(\theta') < \widehat{\mathcal{L}}(\theta) - \bar{\mathcal{L}}(\theta) \le \epsilon/2$. Else, if $\widehat{\pi} < \bar{\pi}$, we have $\epsilon/2 \ge \widehat{\mathcal{L}}(\theta'') - \bar{\mathcal{L}}(\theta'') > \widehat{\mathcal{L}}(\theta') - \bar{\mathcal{L}}(\theta') > \widehat{\mathcal{L}}(\theta) - \bar{\mathcal{L}}(\theta) \ge -\epsilon/2$. Finally, if $\widehat{\pi} = \bar{\pi}$ then $\widehat{\mathcal{L}}(\theta'') - \bar{\mathcal{L}}(\theta'') = \widehat{\mathcal{L}}(\theta') - \bar{\mathcal{L}}(\theta') = \widehat{\mathcal{L}}(\theta) - \bar{\mathcal{L}}(\theta) = 0$. □

The following theorem shows that the expected log-likelihood of the maximum likelihood estimate converges in probability to that of the optimal, as the data size m increases.

Theorem 7.10. *Let $\widehat{\theta} = (\widehat{\mathcal{G}}, \widehat{q})$ be the maximum likelihood estimate in eq.*(7.7) *and $\bar{\theta} = (\bar{\mathcal{G}}, \bar{q})$ be the maximum expected likelihood estimate, i.e. $\widehat{\theta} = \arg\max_{\theta\in\Upsilon}\widehat{\mathcal{L}}(\theta)$ and $\bar{\theta} = \arg\max_{\theta\in\Upsilon}\bar{\mathcal{L}}(\theta)$, then with probability at least $1-\delta$:*

$$\bar{\mathcal{L}}(\widehat{\theta}) \ge \bar{\mathcal{L}}(\bar{\theta}) - \left(\log\max(2m, \tfrac{1}{1-\bar{q}}) + n\log 2\right)\sqrt{\tfrac{2}{m}\left(\log d(\mathcal{H}) + \log\tfrac{4}{\delta}\right)} \qquad (7.10)$$

where $\mathcal{H}$ is the class of games of interest, $\Upsilon = \{(\mathcal{G}, q) \mid \mathcal{G} \in \mathcal{H} \wedge 0 < \pi(\mathcal{G}) < q < 1\}$ is the hypothesis space of non-trivial identifiable games and $d(\mathcal{H}) \equiv |\cup_{\mathcal{G}\in\mathcal{H}}\{\mathcal{NE}(\mathcal{G})\}|$ is the number of all possible games in $\mathcal{H}$ (identified by their Nash equilibria sets).

Proof. First our objective is to find a lower bound for $\mathbb{P}[\bar{\mathcal{L}}(\widehat{\theta}) - \bar{\mathcal{L}}(\bar{\theta}) \ge -\epsilon] \ge \mathbb{P}[\bar{\mathcal{L}}(\widehat{\theta}) - \bar{\mathcal{L}}(\bar{\theta}) \ge -\epsilon + (\widehat{\mathcal{L}}(\widehat{\theta}) - \widehat{\mathcal{L}}(\bar{\theta}))] \ge \mathbb{P}[-\widehat{\mathcal{L}}(\widehat{\theta}) + \bar{\mathcal{L}}(\widehat{\theta}) \ge -\frac{\epsilon}{2}, \widehat{\mathcal{L}}(\bar{\theta}) - \bar{\mathcal{L}}(\bar{\theta}) \ge -\frac{\epsilon}{2}] = \mathbb{P}[\widehat{\mathcal{L}}(\widehat{\theta}) - \bar{\mathcal{L}}(\widehat{\theta}) \le \frac{\epsilon}{2}, \widehat{\mathcal{L}}(\bar{\theta}) - \bar{\mathcal{L}}(\bar{\theta}) \ge -\frac{\epsilon}{2}] = 1 - \mathbb{P}[\widehat{\mathcal{L}}(\widehat{\theta}) - \bar{\mathcal{L}}(\widehat{\theta}) > \frac{\epsilon}{2} \vee \widehat{\mathcal{L}}(\bar{\theta}) - \bar{\mathcal{L}}(\bar{\theta}) < -\frac{\epsilon}{2}]$.

Let $\widetilde{q} \equiv \max(1 - \frac{1}{2m}, \bar{q})$. Now, we have $\mathbb{P}[\widehat{\mathcal{L}}(\widehat{\theta}) - \bar{\mathcal{L}}(\widehat{\theta}) > \frac{\epsilon}{2} \vee \widehat{\mathcal{L}}(\bar{\theta}) - \bar{\mathcal{L}}(\bar{\theta}) < -\frac{\epsilon}{2}] \le \mathbb{P}[(\exists\theta \in \Upsilon, q \le \widetilde{q})\ |\widehat{\mathcal{L}}(\theta) - \bar{\mathcal{L}}(\theta)| > \frac{\epsilon}{2}] = \mathbb{P}[(\exists\theta, \mathcal{G} \in \mathcal{H}, q \in \{\pi(\mathcal{G}), \widetilde{q}\})\ |\widehat{\mathcal{L}}(\theta) - \bar{\mathcal{L}}(\theta)| > \frac{\epsilon}{2}]$. The last equality follows from invoking Lemma 7.9.

Note that $\mathbb{E}[\widehat{\mathcal{L}}(\theta)] = \bar{\mathcal{L}}(\theta)$ and that since $\pi(\mathcal{G}) \le q \le \widetilde{q}$, the log-likelihood is bounded as $(\forall\mathbf{x})\ -B \le \log p_{(\mathcal{G},q)}(\mathbf{x}) \le 0$, where $B = \log\frac{1}{1-\widetilde{q}} + n\log 2 = \log\max(2m, \frac{1}{1-\bar{q}}) + n\log 2$. Therefore, by Hoeffding's inequality, we have $\mathbb{P}[|\widehat{\mathcal{L}}(\theta) - \bar{\mathcal{L}}(\theta)| > \frac{\epsilon}{2}] \le 2e^{-\frac{m\epsilon^2}{2B^2}}$.

Furthermore, note that there are $2d(\mathcal{H})$ possible parameters θ, since we need to consider only two values of $q \in \{\pi(\mathcal{G}), \widetilde{q}\})$ and because the number of all possible games in $\mathcal{H}$ (identified by their Nash equilibria sets) is $d(\mathcal{H}) \equiv |\cup_{\mathcal{G}\in\mathcal{H}}\{\mathcal{NE}(\mathcal{G})\}|$. Therefore, by the union bound we get the following uniform convergence $\mathbb{P}[(\exists\theta, \mathcal{G} \in \mathcal{H}, q \in \{\pi(\mathcal{G}), \widetilde{q}\})\ |\widehat{\mathcal{L}}(\theta) - \bar{\mathcal{L}}(\theta)| > \frac{\epsilon}{2}] \le 4d(\mathcal{H})\mathbb{P}[|\widehat{\mathcal{L}}(\theta) - \bar{\mathcal{L}}(\theta)| > \frac{\epsilon}{2}] \le 4d(\mathcal{H})e^{-\frac{m\epsilon^2}{2B^2}} = \delta$. Finally, by solving for δ we prove our claim. □

The following theorem establishes the complexity of the class of linear influence games, which implies that the term $\log d(\mathcal{H})$ of the generalization bound in Theorem 7.10 is only polynomial in the number of players n.

Theorem 7.11. *Let $\mathcal{H}$ be the class of linear influence games. Then $d(\mathcal{H}) \equiv |\cup_{\mathcal{G}\in\mathcal{H}}\{\mathcal{NE}(\mathcal{G})\}| \le 2^{n\frac{n(n+1)}{2}+1} \le 2^{n^3}$.*

Proof. The logarithm of the number of possible pure-strategy Nash equilibria sets supported by $\mathcal{H}$ (i.e., that can be produced by some game in $\mathcal{H}$) is upper bounded by the VC-dimension of the class of neural networks with a single hidden layer of n units and $n + \binom{n}{2}$ input units, linear threshold activation functions, and constant output weights.

For every linear influence game $\mathcal{G} = (\mathbf{W}, \mathbf{b})$ in $\mathcal{H}$, define the neural network with a single layer of n hidden units, n of the inputs corresponds to the linear terms $x_1, \dots, x_n$ and $\binom{n}{2}$ corresponds to the quadratic polynomial terms $x_i x_j$ for all pairs of players (i, j), $1 \leq i < j \leq n$. For every hidden unit i, the weights corresponding to the linear terms $x_1, \dots, x_n$ are $-b_1, \dots, -b_n$, respectively, while the weights corresponding to the quadratic terms $x_i x_j$ are $-w_{ij}$, for all pairs of players (i, j), $1 \leq i < j \leq n$, respectively. The weights of the bias term of all the hidden units are set to 0. All n output weights are set to 1 while the weight of the output bias term is set to 0. The output of the neural network is $1[\mathbf{x} \notin \mathcal{NE}(\mathcal{G})]$. Note that we define the neural network to classify non-equilibrium as opposed to equilibrium to keep the convention in the neural network literature to define the threshold function to output 0 for input 0. The alternative is to redefine the threshold function to output 1 instead for input 0.

Finally, we use the VC-dimension of neural networks [Sontag, 1998]. □

From Theorems 7.10 and 7.11, we state the generalization bounds for linear influence games.

Corollary 7.12. *Let $\widehat{\theta} = (\widehat{\mathcal{G}}, \widehat{q})$ be the maximum likelihood estimate in eq.(7.7) and $\bar{\theta} = (\bar{\mathcal{G}}, \bar{q})$ be the maximum expected likelihood estimate, i.e. $\widehat{\theta} = \arg\max_{\theta \in \Upsilon} \widehat{\mathcal{L}}(\theta)$ and $\bar{\theta} = \arg\max_{\theta \in \Upsilon} \bar{\mathcal{L}}(\theta)$, then with probability at least $1 - \delta$:*

$$\bar{\mathcal{L}}(\widehat{\theta}) \geq \bar{\mathcal{L}}(\bar{\theta}) - \left(\log \max(2m, \tfrac{1}{1-\bar{q}}) + n \log 2\right) \sqrt{\tfrac{2}{m}\left(n^3 \log 2 + \log \tfrac{4}{\delta}\right)} \qquad (7.11)$$

where $\mathcal{H}$ is the class of linear influence games and $\Upsilon = \{(\mathcal{G}, q) \mid \mathcal{G} \in \mathcal{H} \wedge 0 < \pi(\mathcal{G}) < q < 1\}$ is the hypothesis space of non-trivial identifiable linear influence games.

7.6 Algorithms

In this section, we approximate the maximum likelihood problem problem by maximizing the number of observed equilibria in the data, suitable for a hypothesis space of games with small true proportion of equilibria. We then present our convex loss minimization approach. We also discuss baseline methods such as sigmoidal approximation and exhaustive search.

First, we discuss some negative results that justifies the use of simple approaches. Counting the number of Nash equilibria is NP-hard for influence games, and so is computing the log-likelihood function and therefore maximum likelihood estimation. This is not a disadvantage relative to probabilistic graphical models, since computing the log-likelihood function is also NP-hard for Ising models and Markov random fields in general, while learning is also NP-hard for Bayesian networks. General approximation techniques such as pseudo-likelihood estimation do not lead to tractable methods for learning linear influence games. From an optimization perspective, the log-likelihood function is not continuous because of the number of equilibria. Therefore, we cannot rely on concepts such as Lipschitz continuity.

Furthermore, bounding the number of equilibria by known bounds for Ising models leads to trivial bounds. (Formal proofs and discussion are included in Appendix K.)

7.6.1 An Exact Quasi-Linear Method for General Games: Sample-Picking

As a first approach, consider solving the maximum likelihood estimation problem in eq.(7.7) by an exact exhaustive search algorithm. This algorithm iterates through all possible Nash equilibria sets, i.e. for $s = 0, \dots, 2^n$, we generate all possible sets of size s with elements from the joint-action space $\{-1, +1\}^n$. Recall that there exists $\binom{2^n}{s}$ of such sets of size s and since $\sum_{s=0}^{2^n} \binom{2^n}{s} = 2^{2^n}$ the search space is super-exponential in the number of players n.

Based on few observations, we can obtain an $\mathcal{O}(m \log m)$ algorithm for m samples. First, note that the above method does not constrain the set of Nash equilibria in any fashion. Therefore, only joint-actions that are observed in the data are candidates of being Nash equilibria in order to maximize the log-likelihood. This is because the introduction of an unobserved joint-action will increase the true proportion of equilibria without increasing the empirical proportion of equilibria and thus leading to a lower log-likelihood in eq.(7.7). Second, given a fixed number of Nash equilibria k, the best strategy would be to pick the k joint-actions that appear more frequently in the observed data. This will maximize the empirical proportion of equilibria, which will maximize the log-likelihood. Based on these observations, we propose Algorithm 7.1.

Algorithm 7.1 Sample-Picking for General Games.

```
Input: Dataset $\mathcal{D} = \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}$
Compute the unique samples $\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(U)}$ and their frequency $\widehat{p}^{(1)}, \dots, \widehat{p}^{(U)}$ in the dataset $\mathcal{D}$
Sort joint-actions by their frequency such that $\widehat{p}^{(1)} \geq \widehat{p}^{(2)} \geq \cdots \geq \widehat{p}^{(U)}$
for each unique sample $k = 1, \dots, U$ do
  Define $\mathcal{G}_k$ by the Nash equilibria set $\mathcal{NE}(\mathcal{G}_k) = \{\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(k)}\}$
  Compute the log-likelihood $\widehat{\mathcal{L}}(\mathcal{G}_k, \widehat{q}_k)$ in eq.(7.7) (note that $\widehat{q}_k = \widehat{\pi}(\mathcal{G}) = \frac{1}{m}(\widehat{p}^{(1)} + \cdots + \widehat{p}^{(k)})$, $\pi(\mathcal{G}) = \frac{k}{2^n}$)
end for
Output: The game $\mathcal{G}_{\widehat{k}}$ such that $\widehat{k} = \arg\max_k \widehat{\mathcal{L}}(\mathcal{G}_k, \widehat{q}_k)$
```

As an aside note, the fact that general games do not constrain the set of Nash equilibria, makes the method more likely to over-fit. On the other hand, influence games will potentially include unobserved equilibria given the linearity constraints in the search space, and thus they would be less likely to over-fit.

7.6.2 An Exact Super-Exponential Method for Influence Games: Exhaustive Search

Note that in the previous subsection, we search in the space of all possible games, not only the linear influence games. First note that *sample-picking* for linear games is NP-hard, i.e. at any iteration of *sample-picking*, checking whether the set of Nash equilibria $\mathcal{NE}$ corresponds to an influence game or not is equivalent to the following constraint satisfaction problem with linear constraints:

$$
\begin{aligned}
\min_{\mathbf{W},\mathbf{b}}\ & 1 \\
\text{s.t. }\ & (\forall \mathbf{x} \in \mathcal{NE})\ x_1({\mathbf{w}_{1,-1}}^{\mathrm{T}}\mathbf{x}_{-1} - b_1) \geq 0 \land \cdots \land x_n({\mathbf{w}_{n,-n}}^{\mathrm{T}}\mathbf{x}_{-n} - b_n) \geq 0 \\
& (\forall \mathbf{x} \notin \mathcal{NE})\ x_1({\mathbf{w}_{1,-1}}^{\mathrm{T}}\mathbf{x}_{-1} - b_1) < 0 \lor \cdots \lor x_n({\mathbf{w}_{n,-n}}^{\mathrm{T}}\mathbf{x}_{-n} - b_n) < 0
\end{aligned}
\tag{7.12}
$$

Note that eq.(7.12) contains "or" operators in order to account for the non-equilibria. This makes the problem of finding the $(\mathbf{W}, \mathbf{b})$ that satisfies such conditions NP-hard for a non-empty complement set $\{-1,+1\}^n - \mathcal{NE}$. Furthermore, since *sample-picking* only consider observed equilibria, the search is not optimal with respect to the space of influence games.

Regarding a more refined approach for enumerating influence games only, note that in an influence game each player separates hypercube vertices with a linear function, i.e. for $\mathbf{v} \equiv (\mathbf{w}_{i,-i}, b_i)$ and $\mathbf{y} \equiv (x_i\mathbf{x}_{-i}, -x_i) \in \{-1,+1\}^n$ we have $x_i({\mathbf{w}_{i,-i}}^{\mathrm{T}}\mathbf{x}_{-i} - b_i) = \mathbf{v}^{\mathrm{T}}\mathbf{y}$. Assume we assign a binary label to each vertex $\mathbf{y}$, then note that not all possible labelings are linearly separable. Labelings which are linearly separable are called *linear threshold functions (LTFs)*. A lower bound of the number of LTFs was first provided in Muroga [1965], which showed that the number of LTFs is at least $\alpha(n) \equiv 2^{0.33048n^2}$. Tighter lower bounds were shown later in Yamija and Ibaraki [1965] for $n \geq 6$ and in Muroga and Toda [1966] for $n \geq 8$. Regarding an upper bound, Winder [1960] showed that the number of LTFs is at most $\beta(n) \equiv 2^{n^2}$. By using such bounds for all players, we can conclude that there is at least $\alpha(n)^n = 2^{0.33048n^3}$ and at most $\beta(n)^n = 2^{n^3}$ influence games (which is indeed another upper bound of the VC-dimension of the class of influence games; the bound in Theorem 7.11 is tighter and uses bounds of the VC-dimension of neural networks). The bounds discussed above would bound the time-complexity of a search algorithm if we could easily enumerate all LTFs for a single player. Unfortunately, this seems to be far from a trivial problem. By using results in Muroga [1971], a weight vector $\mathbf{v}$ with integer entries such that $(\forall i)\ |v_i| \leq \beta(n) \equiv (n+1)^{(n+1)/2}/2^n$ is sufficient to realize all possible LTFs. Therefore we can conclude that enumerating influence games takes at most $(2\beta(n)+1)^{n^2} \approx (\frac{\sqrt{n+1}}{2})^{n^3}$ steps, and we propose the use of this method only for $n \leq 4$.

For $n = 4$ we found that the number of influence games is 23,706. Experimentally, we did not find differences between this method and *sample-picking* since most of the time, the model with maximum likelihood was an influence game.

7.6.3 From Maximum Likelihood to Maximum Empirical Proportion of Equilibria

We approximately perform maximum likelihood estimation for influence games, by maximizing the *empirical proportion of equilibria*, i.e. the equilibria in the observed data. This strategy allows us to avoid computing $\pi(\mathcal{G})$ as in eq.(7.4) for maximum likelihood estimation (given its dependence on $|\mathcal{NE}(\mathcal{G})|$). We propose this approach for games with small true proportion of equilibria with high probability, i.e. with probability at least $1-\delta$, we have $\pi(\mathcal{G}) \leq \frac{\kappa^n}{\delta}$ for $0 < \kappa < 1$. Particularly, we will show in Section 7.7 that for influence games we have $\kappa = 3/4$. Given this, our approximate problem relies on a bound of the

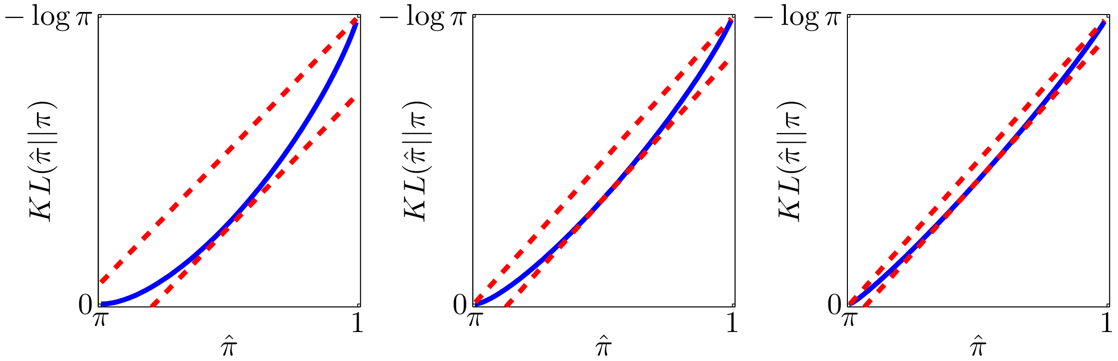


Figure 7.2: KL divergence (blue) and bounds derived in Lemma 7.13 (red) for $\pi = (3/4)^n$ where $n = 9$ (left), $n = 18$ (center) and $n = 36$ (right). Note that the bounds are very informative when $n \to +\infty$ (or equivalently when $\pi \to 0$).

log-likelihood that holds with high probability. We also show that under very mild conditions, the parameters $(\mathcal{G}, q)$ belong to the hypothesis space of the original problem with high probability.

First, we derive bounds on the log-likelihood function.

Lemma 7.13. *Given a non-trivial game $\mathcal{G}$ with $0 < \pi(\mathcal{G}) < \widehat{\pi}(\mathcal{G})$, the KL divergence in the log-likelihood function in eq.(7.7) is bounded as follows:*

$$-\widehat{\pi}(\mathcal{G}) \log \pi(\mathcal{G}) - \log 2 < KL(\widehat{\pi}(\mathcal{G}) \| \pi(\mathcal{G})) < -\widehat{\pi}(\mathcal{G}) \log \pi(\mathcal{G}) \tag{7.13}$$

Proof. Let $\pi \equiv \pi(\mathcal{G})$ and $\widehat{\pi} \equiv \widehat{\pi}(\mathcal{G})$. Note that $\alpha(\pi) \equiv \lim_{\widehat{\pi} \to 0} KL(\widehat{\pi} \| \pi) = 0$ and $\beta(\pi) \equiv \lim_{\widehat{\pi} \to 1} KL(\widehat{\pi} \| \pi) = -\log \pi \leq n \log 2$. Since the function is convex we can upper-bound it by $\alpha(\pi) + (\beta(\pi) - \alpha(\pi))\widehat{\pi} = -\widehat{\pi} \log \pi$.

To find a lower bound, we find the point in which the derivative of the original function is equal to the slope of the upper bound, i.e. $\frac{\partial KL(\widehat{\pi}\|\pi)}{\partial \widehat{\pi}} = \beta(\pi) - \alpha(\pi) = -\log \pi$, which gives $\widehat{\pi}^* = \frac{1}{2-\pi}$. Then, the maximum difference between the upper bound and the original function is given by $\lim_{\pi \to 0} -\widehat{\pi}^* \log \pi - KL(\widehat{\pi}^* \| \pi) = \log 2$. □

Note that the lower and upper bounds are very informative when $\pi(\mathcal{G}) \to 0$ (or in our setting when $n \to +\infty$), since $\log 2$ becomes small when compared to $-\log \pi(\mathcal{G})$, as shown in Figure 7.2.

Next, we derive the problem of maximizing the empirical proportion of equilibria from the maximum likelihood estimation problem.

Theorem 7.14. *Assume that with probability at least $1-\delta$ we have $\pi(\mathcal{G}) \leq \frac{\kappa^n}{\delta}$ for $0 < \kappa < 1$. Maximizing a lower bound (with high probability) of the log-likelihood in eq.(7.7) is equivalent to maximizing the empirical proportion of equilibria:*

$$\max_{\mathcal{G} \in \mathcal{H}} \ \widehat{\pi}(\mathcal{G}) \tag{7.14}$$

furthermore, for all games $\mathcal{G}$ such that $\widehat{\pi}(\mathcal{G}) \geq \gamma$ for some $0 < \gamma < 1/2$, for sufficiently large $n > \log_\kappa(\delta\gamma)$ and optimal mixture parameter $\widehat{q} = \min(\widehat{\pi}(\mathcal{G}), 1 - \frac{1}{2m})$, we have $(\mathcal{G}, \widehat{q}) \in \Upsilon$, where $\Upsilon = \{(\mathcal{G}, q) \mid \mathcal{G} \in \mathcal{H} \wedge 0 < \pi(\mathcal{G}) < q < 1\}$ is the hypothesis space of non-trivial identifiable games.

Proof. By applying the lower bound in Lemma 7.13 in eq.(7.7) to non-trivial games, we have $\widehat{\mathcal{L}}(\mathcal{G},\widehat{q}) = KL(\widehat{\pi}(\mathcal{G})\|\pi(\mathcal{G})) - KL(\widehat{\pi}(\mathcal{G})\|\widehat{q}) - n\log 2 > -\widehat{\pi}(\mathcal{G})\log\pi(\mathcal{G}) - KL(\widehat{\pi}(\mathcal{G})\|\widehat{q}) - (n+1)\log 2$. Since $\pi(\mathcal{G}) \leq \frac{\kappa^n}{\delta}$, we have $-\log\pi(\mathcal{G}) \geq -\log\frac{\kappa^n}{\delta}$. Therefore $\widehat{\mathcal{L}}(\mathcal{G},\widehat{q}) > -\widehat{\pi}(\mathcal{G})\log\frac{\kappa^n}{\delta} - KL(\widehat{\pi}(\mathcal{G})\|\widehat{q}) - (n+1)\log 2$. Regarding the term $KL(\widehat{\pi}(\mathcal{G})\|\widehat{q})$, if $\widehat{\pi}(\mathcal{G}) < 1 \Rightarrow KL(\widehat{\pi}(\mathcal{G})\|\widehat{q}) = KL(\widehat{\pi}(\mathcal{G})\|\widehat{\pi}(\mathcal{G})) = 0$, and if $\widehat{\pi}(\mathcal{G}) = 1 \Rightarrow KL(\widehat{\pi}(\mathcal{G})\|\widehat{q}) = KL(1\|1-\frac{1}{2m}) = -\log(1-\frac{1}{2m}) \leq \log 2$ and approaches 0 when $m \to +\infty$. Maximizing the lower bound of the log-likelihood becomes $\max_{\mathcal{G}\in\mathcal{H}}\widehat{\pi}(\mathcal{G})$ by removing the constant terms that do not depend on $\mathcal{G}$.

In order to prove $(\mathcal{G},\widehat{q}) \in \Upsilon$ we need to prove $0 < \pi(\mathcal{G}) < \widehat{q} < 1$. For proving the first inequality $0 < \pi(\mathcal{G})$, note that $\widehat{\pi}(\mathcal{G}) \geq \gamma > 0$, and therefore $\mathcal{G}$ has at least one equilibria. For proving the third inequality $\widehat{q} < 1$, note that $\widehat{q} = \min(\widehat{\pi}(\mathcal{G}), 1-\frac{1}{2m}) < 1$. For proving the second inequality $\pi(\mathcal{G}) < \widehat{q}$, we need to prove $\pi(\mathcal{G}) < \widehat{\pi}(\mathcal{G})$ and $\pi(\mathcal{G}) < 1-\frac{1}{2m}$. Since $\pi(\mathcal{G}) \leq \frac{\kappa^n}{\delta}$ and $\gamma \leq \widehat{\pi}(\mathcal{G})$, it suffices to prove $\frac{(3/4)^n}{\delta} < \gamma \Rightarrow \pi(\mathcal{G}) < \widehat{\pi}(\mathcal{G})$. Similarly we need to prove $\frac{(3/4)^n}{\delta} < 1-\frac{1}{2m} \Rightarrow \pi(\mathcal{G}) < 1-\frac{1}{2m}$. Putting both together, we have $\frac{(3/4)^n}{\delta} < \min(\gamma, 1-\frac{1}{2m}) = \gamma$ since $\gamma < 1/2$ and $1-\frac{1}{2m} \geq 1/2$. Finally, $\frac{(3/4)^n}{\delta} < \gamma \Leftrightarrow n > \log_\kappa(\delta\gamma)$. □

7.6.4 A Non-Concave Maximization Method: Sigmoidal Approximation

A very simple optimization approach can be devised by using a sigmoid in order to approximate the 0/1 function $1[z \geq 0]$ in the maximum likelihood problem of eq.(7.7) as well as when maximizing the empirical proportion of equilibria as in eq.(7.14). We use the following sigmoidal approximation:

$$1[z \geq 0] \approx H_{\alpha,\beta}(z) \equiv \tfrac{1}{2}(1 + \tanh(\tfrac{z}{\beta} - \mathrm{arctanh}(1 - 2\alpha^{1/n}))) \tag{7.15}$$

The additional term α ensures that for $\mathcal{G} = (\mathbf{W},\mathbf{b}), \mathbf{W} = \mathbf{0}, \mathbf{b} = \mathbf{0}$ we get $1[\mathbf{x} \in \mathcal{NE}(\mathcal{G})] \approx H_{\alpha,\beta}(0)^n = \alpha$. We perform gradient ascent on these objective functions that have many local maxima. Note that when maximizing the "sigmoidal" likelihood, each step of the gradient ascent is NP-hard due to the "sigmoidal" true proportion of equilibria. Therefore, we propose the use of the sigmoidal maximum likelihood only for $n \leq 15$.

In our implementation, we add an ℓ_1-norm regularizer $-\rho\|\mathbf{W}\|_1$ where $\rho > 0$ to both maximization problems. The ℓ_1-norm regularizer encourages sparseness and attempts to lower the generalization error by controlling over-fitting.

7.6.5 Our Proposed Approach: Convex Loss Minimization

From an optimization perspective, it is more convenient to minimize a convex objective instead of a sigmoidal approximation in order to avoid the many local minima.

Note that maximizing the empirical proportion of equilibria in eq.(7.14) is equivalent to minimizing the empirical proportion of non-equilibria, i.e. $\min_{\mathcal{G}\in\mathcal{H}}(1 - \widehat{\pi}(\mathcal{G}))$. Furthermore, $1 - \widehat{\pi}(\mathcal{G}) = \frac{1}{m}\sum_l 1[\mathbf{x}^{(l)} \notin \mathcal{NE}(\mathcal{G})]$. Denote by ℓ the 0/1 loss, i.e. $\ell(z) = 1[z < 0]$. For influence games, maximizing the empirical proportion of equilibria in eq.(7.14) is equivalent to solving the loss minimization problem:

$$\min_{\mathbf{W},\mathbf{b}} \frac{1}{m}\sum_l \max_i \ell(x_i^{(l)}(\mathbf{w}_{i,-i}{}^{\mathrm{T}}\mathbf{x}_{-i}^{(l)} - b_i)) \tag{7.16}$$

We can further relax this problem by introducing convex upper bounds of the 0/1 loss. Note that the use of convex losses also avoids the trivial solution of eq.(7.16), i.e. $\mathbf{W} = \mathbf{0}, \mathbf{b} = \mathbf{0}$ (which obtains the lowest log-likelihood as discussed in Remark 7.8). Intuitively speaking, note that minimizing the logistic loss $\ell(z) = \log(1 + e^{-z})$ will make $z \to +\infty$, while minimizing the hinge loss $\ell(z) = \max(0, 1 - z)$ will make $z \to 1$ unlike the 0/1 loss $\ell(z) = 1[z < 0]$ that only requires $z = 0$ in order to be minimized. In what follows, we develop four efficient methods for solving eq.(7.16) under specific choices of loss functions, i.e. hinge and logistic.

In our implementation, we add an ℓ_1-norm regularizer $\rho\|\mathbf{W}\|_1$ where $\rho > 0$ to all the minimization problems. The ℓ_1-norm regularizer encourages sparseness and attempts to lower the generalization error by controlling over-fitting.

Independent Support Vector Machines and Logistic Regression. We can relax the loss minimization problem in eq.(7.16) by using the loose bound $\max_i \ell(z_i) \leq \sum_i \ell(z_i)$. This relaxation simplifies the original problem into several independent problems. For each player i, we train the weights $(\mathbf{w}_{i,-i}, b_i)$ in order to predict independent (disjoint) actions. This leads to *1-norm SVMs* of Bradley and Mangasarian [1998], Zhu et al. [2003] and ℓ_1-regularized logistic regression. We solve the latter with the *ℓ_1-projection method* of Schmidt et al. [2007a]. While the training is independent, our goal is not the prediction for independent players but the characterization of joint-actions. The use of these well known techniques in our context is novel, since we interpret the output of SVMs and logistic regression as the parameters of an influence game. Therefore, we use the parameters to measure empirical and true proportion of equilibria, KL divergence and log-likelihood in our probabilistic model.

Simultaneous Support Vector Machines. While converting the loss minimization problem in eq.(7.16) by using loose bounds allow to obtain several independent problems with small number of variables, a second reasonable strategy would be to use tighter bounds at the expense of obtaining a single optimization problem with a higher number of variables.

For the hinge loss $\ell(z) = \max(0, 1 - z)$, we have $\max_i \ell(z_i) = \max(0, 1 - z_1, \dots, 1 - z_n)$ and the loss minimization problem in eq.(7.16) becomes the following primal linear program:

$$\begin{aligned} &\min_{\mathbf{W},\mathbf{b},\boldsymbol{\xi}} \; \frac{1}{m}\sum_l \xi_l + \rho\|\mathbf{W}\|_1 \\ &\text{s.t. } (\forall l, i)\; x_i^{(l)}(\mathbf{w}_{i,-i}{}^{\mathrm{T}}\mathbf{x}_{-i}^{(l)} - b_i) \geq 1 - \xi_l \;\;,\;\; (\forall l)\; \xi_l \geq 0 \end{aligned} \tag{7.17}$$

where $\rho > 0$.

Note that eq.(7.17) is equivalent to a linear program since we can set $\mathbf{W} = \mathbf{W}^+ - \mathbf{W}^-$, $\|\mathbf{W}\|_1 = \sum_{ij} w_{ij}^+ + w_{ij}^-$ and add the constraints $\mathbf{W}^+ \geq \mathbf{0}$ and $\mathbf{W}^- \geq \mathbf{0}$. We follow the regular SVM derivation by adding slack variables ξ_l for each sample l. This problem is a generalization of *1-norm SVMs* of Bradley and Mangasarian [1998], Zhu et al. [2003].

By Lagrangian duality, the dual of the problem in eq.(7.17) is the following linear program:

$$\begin{array}{ll} \max_{\boldsymbol{\alpha}} & \sum_{li} \alpha_{li} \\ \text{s.t.} & (\forall i)\ \| \textstyle\sum_l \alpha_{li} x_i^{(l)} \mathbf{x}_{-i}^{(l)} \|_\infty \leq \rho \quad , \quad (\forall l, i)\ \alpha_{li} \geq 0 \\ & (\forall i)\ \textstyle\sum_l \alpha_{li} x_i^{(l)} = 0 \quad , \qquad\qquad (\forall l)\ \textstyle\sum_i \alpha_{li} \leq \frac{1}{m} \end{array} \tag{7.18}$$

Furthermore, strong duality holds in this case. Note that eq.(7.18) is equivalent to a linear program since we can transform the constraint $\|\mathbf{c}\|_\infty \leq \rho$ into $-\rho\mathbf{1} \leq \mathbf{c} \leq \rho\mathbf{1}$.

Simultaneous Logistic Regression. For the logistic loss $\ell(z) = \log(1 + e^{-z})$, we could use the non-smooth loss $\max_i \ell(z_i)$ directly. Instead, we chose a smooth upper bound, i.e. $\log(1 + \sum_i e^{-z_i})$ (Discussion is included in Appendix L.) The loss minimization problem in eq.(7.16) becomes:

$$\min_{\mathbf{W},\mathbf{b}} \frac{1}{m} \sum_l \log(1 + \textstyle\sum_i e^{-x_i^{(l)} (\mathbf{w}_{i,-i}{}^{\mathrm{T}} \mathbf{x}_{-i}^{(l)} - b_i)}) + \rho \|\mathbf{W}\|_1 \tag{7.19}$$

where $\rho > 0$.

In our implementation, we use the *ℓ_1-projection method* of Schmidt et al. [2007a] for optimizing eq.(7.19). This method performs a *limited memory Broyden-Fletcher-Goldfarb-Shanno* (L-BFGS) step in an expanded model (i.e. $\mathbf{W} = \mathbf{W}^+ - \mathbf{W}^-$, $\|\mathbf{W}\|_1 = \sum_{ij} w_{ij}^+ + w_{ij}^-$) followed by a projection onto the non-negative orthant to enforce $\mathbf{W}^+ \geq \mathbf{0}$ and $\mathbf{W}^- \geq \mathbf{0}$.

7.7 True Proportion of Equilibria

In this section, we justify the use of convex loss minimization for learning the structure and parameters of influence games. We define *absolute indifference* of players and show that our convex loss minimization approach produces games in which all players are non-absolutely-indifferent. We then provide a bound of the true proportion of equilibria with high probability. Our bound only assumes independence of weight vectors among players. Our bound is distribution-free, i.e. we do not assume a specific distribution for the weight vector of each player. Furthermore, we do not assume any connectivity properties of the underlying graph.

Parallel to our analysis, Daskalakis et al. [2011] analyzed a different setting: random games which structure is drawn from the Erdős-Rényi model (i.e. each edge is present independently with the same probability p) and utility functions which are random tables. The analysis in Daskalakis et al. [2011], while more general than ours (which only focus on influence games), it is at the same time more restricted since it assumes either the Erdős-Rényi model for random structures or connectivity properties for deterministic structures.

7.7.1 Convex Loss Minimization Produces Non-Absolutely- Indifferent Players

First, we define the notion of *absolute indifference* of players. Our goal in this subsection is to show that our proposed convex loss algorithms produce influence games in which all players are non-absolutely-indifferent and therefore every player defines constraints to the true proportion of equilibria.

Definition 7.15. *Given an influence game* $\mathcal{G} = (\mathbf{W}, \mathbf{b})$*, we say a player* i *is* absolutely indifferent *if and only if* $(\mathbf{w}_{i,-i}, b_i) = \mathbf{0}$*, and* non-absolutely-indifferent *if and only if* $(\mathbf{w}_{i,-i}, b_i) \neq \mathbf{0}$.

Next, we concentrate on the first ingredient for our bound of the true proportion of equilibria. We show that independent and simultaneous SVM and logistic regression produce games in which all players are non-absolutely-indifferent except for some "degenerate" cases. The following lemma applies to independent SVMs for $c^{(l)} = 0$ and simultaneous SVMs for $c^{(l)} = \max(0, \max_{j \neq i} (1 - x_j^{(l)}(\mathbf{w}_{i,-i}{}^{\mathrm{T}}\mathbf{x}_{-i}^{(l)} - b_i)))$.

Lemma 7.16. *Given* $(\forall l)\ c^{(l)} \geq 0$*, the minimization of the hinge training loss* $\widehat{\ell}(\mathbf{w}_{i,-i}, b_i) = \frac{1}{m}\sum_l \max(c^{(l)}, 1 - x_i^{(l)}(\mathbf{w}_{i,-i}{}^{\mathrm{T}}\mathbf{x}_{-i}^{(l)} - b_i))$ *guarantees non-absolutely-indifference of player* i *except for some "degenerate" cases, i.e. the optimal solution* $(\mathbf{w}_{i,-i}^*, b_i^*) = \mathbf{0}$ *if and only if* $(\forall j \neq i)\ \sum_l 1[x_i^{(l)}x_j^{(l)}{=}1]u^{(l)} = \sum_l 1[x_i^{(l)}x_j^{(l)}{=}{-1}]u^{(l)}$ *and* $\sum_l 1[x_i^{(l)}{=}1]u^{(l)} = \sum_l 1[x_i^{(l)}{=}{-1}]u^{(l)}$ *where* $u^{(l)}$ *is defined as* $c^{(l)} > 1 \Leftrightarrow u^{(l)} = 0$*,* $c^{(l)} < 1 \Leftrightarrow u^{(l)} = 1$ *and* $c^{(l)} = 1 \Leftrightarrow u^{(l)} \in [0;1]$.

Proof. Let $f_i(\mathbf{x}_{-i}) \equiv \mathbf{w}_{i,-i}{}^{\mathrm{T}}\mathbf{x}_{-i} - b_i$. By noting that $\max(\alpha, \beta) = \max_{0 \leq u \leq 1} (\alpha + u(\beta - \alpha))$, we can rewrite $\widehat{\ell}(\mathbf{w}_{i,-i}, b_i) = \frac{1}{m}\sum_l \max_{0 \leq u^{(l)} \leq 1} (c^{(l)} + u^{(l)}(1 - x_i^{(l)}f_i(\mathbf{x}_{-i}^{(l)}) - c^{(l)}))$.

Note that $\widehat{\ell}$ has the minimizer $(\mathbf{w}_{i,-i}^*, b_i^*) = \mathbf{0}$ if and only if $\mathbf{0}$ belongs to the subdifferential set of the non-smooth function $\widehat{\ell}$ at $(\mathbf{w}_{i,-i}, b_i) = \mathbf{0}$. In order to maximize $\widehat{\ell}$, we have $c^{(l)} > 1 - x_i^{(l)}f_i(\mathbf{x}_{-i}^{(l)}) \Leftrightarrow u^{(l)} = 0$, $c^{(l)} < 1 - x_i^{(l)}f_i(\mathbf{x}_{-i}^{(l)}) \Leftrightarrow u^{(l)} = 1$ and $c^{(l)} = 1 - x_i^{(l)}f_i(\mathbf{x}_{-i}^{(l)}) \Leftrightarrow u^{(l)} \in [0;1]$. The previous rules simplify at the solution under analysis, since $(\mathbf{w}_{i,-i}, b_i) = \mathbf{0} \Rightarrow f_i(\mathbf{x}_{-i}^{(l)}) = 0$.

Let $g_j(\mathbf{w}_{i,-i}, b_i) \equiv \frac{\partial \widehat{\ell}}{\partial w_{ij}}(\mathbf{w}_{i,-i}, b_i)$ and $h(\mathbf{w}_{i,-i}, b_i) \equiv \frac{\partial \widehat{\ell}}{\partial b_i}(\mathbf{w}_{i,-i}, b_i)$. By making $(\forall j \neq i)\ 0 \in g_j(\mathbf{0}, 0)$ and $0 \in h(\mathbf{0}, 0)$, we get $(\forall j \neq i)\ \sum_l x_i^{(l)}x_j^{(l)}u^{(l)} = 0$ and $\sum_l x_i^{(l)}u^{(l)} = 0$. Finally, by noting that $x_i^{(l)} \in \{-1, 1\}$, we prove our claim. □

Remark 7.17. *Note that for independent SVMs, the "degenerate" cases in Lemma 7.16 simplify to* $(\forall j \neq i)\ \sum_l 1[x_i^{(l)}x_j^{(l)} = 1] = \frac{m}{2}$ *and* $\sum_l 1[x_i^{(l)} = 1] = \frac{m}{2}$.

The following lemma applies to independent logistic regression for $c^{(l)} = 0$ and simultaneous logistic regression for $c^{(l)} = \sum_{j \neq i} e^{-x_j^{(l)}(\mathbf{w}_{i,-i}{}^{\mathrm{T}}\mathbf{x}_{-i}^{(l)} - b_i)}$.

Lemma 7.18. *Given* $(\forall l)\ c^{(l)} \geq 0$*, the minimization of the logistic training loss* $\widehat{\ell}(\mathbf{w}_{i,-i}, b_i) = \frac{1}{m}\sum_l \log(c^{(l)} + 1 + e^{-x_i^{(l)}(\mathbf{w}_{i,-i}{}^{\mathrm{T}}\mathbf{x}_{-i}^{(l)} - b_i)})$ *guarantees non-absolutely-indifference of player* i *except for some "degenerate" cases, i.e. the optimal solution* $(\mathbf{w}_{i,-i}^*, b_i^*) = \mathbf{0}$ *if and only if* $(\forall j \neq i)\ \sum_l \frac{1[x_i^{(l)}x_j^{(l)}=1]}{c^{(l)}+2} = \sum_l \frac{1[x_i^{(l)}x_j^{(l)}=-1]}{c^{(l)}+2}$ *and* $\sum_l \frac{1[x_i^{(l)}=1]}{c^{(l)}+2} = \sum_l \frac{1[x_i^{(l)}=-1]}{c^{(l)}+2}$.

Proof. Note that $\widehat{\ell}$ has the minimizer $(\mathbf{w}_{i,-i}^*, b_i^*) = \mathbf{0}$ if and only if the gradient of the smooth function $\widehat{\ell}$ is $\mathbf{0}$ at $(\mathbf{w}_{i,-i}, b_i) = \mathbf{0}$. Let $g_j(\mathbf{w}_{i,-i}, b_i) \equiv \frac{\partial \widehat{\ell}}{\partial w_{ij}}(\mathbf{w}_{i,-i}, b_i)$ and $h(\mathbf{w}_{i,-i}, b_i) \equiv \frac{\partial \widehat{\ell}}{\partial b_i}(\mathbf{w}_{i,-i}, b_i)$. By making $(\forall j \neq i)\ g_j(\mathbf{0}, 0) = 0$ and $h(\mathbf{0}, 0) = 0$, we get $(\forall j \neq i)\ \sum_l \frac{x_i^{(l)}x_j^{(l)}}{c^{(l)}+2} = 0$ and $\sum_l \frac{x_i^{(l)}}{c^{(l)}+2} = 0$. Finally, by noting that $x_i^{(l)} \in \{-1, 1\}$, we prove our claim. □

Remark 7.19. *Note that for independent logistic regression, the "degenerate" cases in Lemma 7.18 simplify to* $(\forall j \neq i)$ $\sum_l 1[x_i^{(l)} x_j^{(l)} = 1] = \frac{m}{2}$ *and* $\sum_l 1[x_i^{(l)} = 1] = \frac{m}{2}$.

Based on these results, after termination of our proposed algorithms, we fix cases in which the optimal solution $(\mathbf{w}^*_{i,-i}, b^*_i) = \mathbf{0}$ by setting $b^*_i = 1$ if the action of player i was mostly -1 or $b^*_i = -1$ otherwise. We point out to the careful reader that we did not include the ℓ_1-regularization term in the above proofs since the subdifferential of $\rho\|\mathbf{w}_{i,-i}\|_1$ vanishes at $\mathbf{w}_{i,-i} = 0$, and therefore our proofs still hold.

7.7.2 Bounding the True Proportion of Equilibria

In what follows, we concentrate on the second ingredient for our bound of the true proportion of equilibria. We show that for a game with a single *non-absolutely-indifferent* player, the true proportion of equilibria is bounded by 3/4.

Lemma 7.20. *Given an influence game* $\mathcal{G} = (\mathbf{W}, \mathbf{b})$ *with non-absolutely-indifferent player i and absolutely-indifferent players* $\forall j \neq i$*, the following statements hold:*

$$\begin{array}{l} \text{i. } \mathbf{x} \in \mathcal{NE}(\mathcal{G}) \Leftrightarrow x_i({\mathbf{w}_{i,-i}}^{\mathrm{T}}\mathbf{x}_{-i} - b_i) \geq 0 \\ \text{ii. } |\mathcal{NE}(\mathcal{G})| = 2^{n-1} + \sum_{\mathbf{x}_{-i}} 1[{\mathbf{w}_{i,-i}}^{\mathrm{T}}\mathbf{x}_{-i} - b_i = 0] \\ \text{iii. } \frac{1}{2} \leq \pi(\mathcal{G}) \leq \frac{3}{4} \end{array} \tag{7.20}$$

Proof. Let $f_i(\mathbf{x}_{-i}) \equiv {\mathbf{w}_{i,-i}}^{\mathrm{T}}\mathbf{x}_{-i} - b_i$. For proving Claim i, note that $1[\mathbf{x} \in \mathcal{NE}(\mathcal{G})] = \min_j 1[x_j f_j(\mathbf{x}_{-j}) \geq 0] = 1[x_i f_i(\mathbf{x}_{-i}) \geq 0] \min_{j \neq i} 1[x_j f_j(\mathbf{x}_{-j}) \geq 0]$. Since all players except i are absolutely-indifferent, we have $(\forall j \neq i)$ $(\mathbf{w}_{j,-j}, b_j) = \mathbf{0} \Rightarrow f_j(\mathbf{x}_{-j}) = 0$ which implies that $\min_{j \neq i} 1[x_j f_j(\mathbf{x}_{-j}) \geq 0] = 1$. Therefore, $1[\mathbf{x} \in \mathcal{NE}(\mathcal{G})] = 1[x_i f_i(\mathbf{x}_{-i}) \geq 0]$.

For proving Claim ii, by Claim i we have $|\mathcal{NE}(\mathcal{G})| = \sum_{\mathbf{x}} 1[x_i f_i(\mathbf{x}_{-i}) \geq 0]$. We can rewrite $|\mathcal{NE}(\mathcal{G})| = \sum_{\mathbf{x}} 1[x_i = +1] 1[f_i(\mathbf{x}_{-i}) \geq 0] + \sum_{\mathbf{x}} 1[x_i = -1] 1[f_i(\mathbf{x}_{-i}) \leq 0]$ or equivalently $|\mathcal{NE}(\mathcal{G})| = \sum_{\mathbf{x}_{-i}} 1[f_i(\mathbf{x}_{-i}) \geq 0] + \sum_{\mathbf{x}_{-i}} 1[f_i(\mathbf{x}_{-i}) \leq 0] = 2^{n-1} + \sum_{\mathbf{x}_{-i}} 1[f_i(\mathbf{x}_{-i}) = 0]$.

For proving Claim iii, by eq.(7.4) and Claim ii we have $\pi(\mathcal{G}) = \frac{|\mathcal{NE}(\mathcal{G})|}{2^n} = \frac{1}{2} + \frac{1}{2^n}\alpha(\mathbf{w}_{i,-i}, b_i)$, where $\alpha(\mathbf{w}_{i,-i}, b_i) \equiv \sum_{\mathbf{x}_{-i}} 1[{\mathbf{w}_{i,-i}}^{\mathrm{T}}\mathbf{x}_{-i} - b_i = 0]$. This proves the lower bound $\pi(\mathcal{G}) \geq \frac{1}{2}$. Geometrically speaking, $\alpha(\mathbf{w}_{i,-i}, b_i)$ is the number of vertices of the $(n-1)$-dimensional hypercube that are covered by the hyperplane with normal $\mathbf{w}_{i,-i}$ and bias b_i. Recall that $(\mathbf{w}_{i,-i}, b_i) \neq \mathbf{0}$. If $\mathbf{w}_{i,-i} = \mathbf{0}$ and $b_i \neq 0$ then $\alpha(\mathbf{w}_{i,-i}, b_i) = \sum_{\mathbf{x}_{-i}} 1[b_i = 0] = 0 \Rightarrow \pi(\mathcal{G}) = \frac{1}{2}$. If $\mathbf{w}_{i,-i} \neq \mathbf{0}$ then as noted in Aichholzer and Aurenhammer [1996] a hyperplane with $n-2$ zeros on $\mathbf{w}_{i,-i}$ (i.e. a $(n-2)$-*parallel hyperplane*) covers exactly half of the 2^{n-1} vertices, the maximum possible. Therefore, $\pi(\mathcal{G}) = \frac{1}{2} + \frac{1}{2^n}\alpha(\mathbf{w}_{i,-i}, b_i) \leq \frac{1}{2} + \frac{2^{n-2}}{2^n} = \frac{3}{4}$. □

Next, we present our bound for the true proportion of equilibria of games in which all players are non-absolutely-indifferent.

Theorem 7.21. *If all players are non-absolutely-indifferent and if the rows of an influence game* $\mathcal{G} = (\mathbf{W}, \mathbf{b})$ *are independent (but not necessarily identically distributed) random vectors, i.e. for every player i,* $(\mathbf{w}_{i,-i}, b_i)$ *is independently drawn from an arbitrary distribution* $\mathcal{P}_i$*, then the expected true proportion of equilibria is bounded as follows:*

$$(1/2)^n \leq \mathbb{E}_{\mathcal{P}_1,\ldots,\mathcal{P}_n}[\pi(\mathcal{G})] \leq (3/4)^n \tag{7.21}$$

furthermore, the following high probability statement holds:

$$\mathbb{P}_{\mathcal{P}_1,\dots,\mathcal{P}_n}[\pi(\mathcal{G}) \leq \tfrac{(3/4)^n}{\delta}] \geq 1 - \delta \tag{7.22}$$

Proof. Let $y_i \equiv 1[x_i(\mathbf{w}_{i,-i}{}^{\mathrm{T}}\mathbf{x}_{-i} - b_i) \geq 0]$, $\mathcal{P} \equiv \{\mathcal{P}_1, \dots, \mathcal{P}_n\}$ and $\mathcal{U}$ the uniform distribution for $\mathbf{x} \in \{-1,+1\}^n$. By eq.(7.4), $\mathbb{E}_{\mathcal{P}}[\pi(\mathcal{G})] = \mathbb{E}_{\mathcal{P}}[\frac{1}{2^n}\sum_{\mathbf{x}}\prod_i y_i] = \mathbb{E}_{\mathcal{P}}[\mathbb{E}_{\mathcal{U}}[\prod_i y_i]] = \mathbb{E}_{\mathcal{U}}[\mathbb{E}_{\mathcal{P}}[\prod_i y_i]]$. Note that each y_i is independent since each $(\mathbf{w}_{i,-i}, b_i)$ is independently distributed. Therefore, $\mathbb{E}_{\mathcal{P}}[\pi(\mathcal{G})] = \mathbb{E}_{\mathcal{U}}[\prod_i \mathbb{E}_{\mathcal{P}_i}[y_i]]$. Similarly each $z_i \equiv \mathbb{E}_{\mathcal{P}_i}[y_i]$ is independent since each $(\mathbf{w}_{i,-i}, b_i)$ is independently distributed. Therefore, $\mathbb{E}_{\mathcal{P}}[\pi(\mathcal{G})] = \mathbb{E}_{\mathcal{U}}[\prod_i z_i] = \prod_i \mathbb{E}_{\mathcal{U}}[z_i] = \prod_i \mathbb{E}_{\mathcal{U}}[\mathbb{E}_{\mathcal{P}_i}[y_i]] = \prod_i \mathbb{E}_{\mathcal{P}_i}[\mathbb{E}_{\mathcal{U}}[y_i]]$. Note that $\mathbb{E}_{\mathcal{U}}[y_i]$ is the true proportion of equilibria of an influence game with non-absolutely-indifferent player i and absolutely-indifferent players $\forall j \neq i$, and therefore $1/2 \leq \mathbb{E}_{\mathcal{U}}[y_i] \leq 3/4$ by Claim iii of Lemma 7.20. Finally, we have $\mathbb{E}_{\mathcal{P}}[\pi(\mathcal{G})] \geq \prod_i \mathbb{E}_{\mathcal{P}_i}[1/2] = (1/2)^n$ and similarly $\mathbb{E}_{\mathcal{P}}[\pi(\mathcal{G})] \leq \prod_i \mathbb{E}_{\mathcal{P}_i}[3/4] = (3/4)^n$.

By Markov's inequality, given that $\pi(\mathcal{G}) \geq 0$, we have $\mathbb{P}_{\mathcal{P}}[\pi(\mathcal{G}) \geq c] \leq \frac{\mathbb{E}_{\mathcal{P}}[\pi(\mathcal{G})]}{c} \leq \frac{(3/4)^n}{c}$. For $c = \frac{(3/4)^n}{\delta} \Rightarrow \mathbb{P}_{\mathcal{P}}[\pi(\mathcal{G}) \geq \frac{(3/4)^n}{\delta}] \leq \delta \Rightarrow \mathbb{P}_{\mathcal{P}}[\pi(\mathcal{G}) \leq \frac{(3/4)^n}{\delta}] \geq 1 - \delta$. □

Remark 7.22. *Under the same assumptions of Theorem 7.21, it is possible to prove that with probability at least* $1 - \delta$ *we have* $\pi(\mathcal{G}) \leq (3/4)^n + 3/8\sqrt{2\log\frac{1}{\delta}}$ *by using Hoeffding's lemma. We point out that such a bound is not better than the Markov's bound derived above.*

7.8 Experimental Results

For learning influence games we used our convex loss methods: independent and simultaneous SVM and logistic regression. Additionally, we used the (super-exponential) exhaustive search method only for $n \leq 4$. As a baseline, we used the sigmoidal maximum likelihood (NP-hard) only for $n \leq 15$ as well as the sigmoidal maximum empirical proportion of equilibria. Regarding the parameters α and β our sigmoidal function in eq.(7.15), we found experimentally that $\alpha = 0.1$ and $\beta = 0.001$ achieved the best results.

We compare learning influence games to learning Ising models. For $n \leq 15$ players, we perform exact ℓ_1-regularized maximum likelihood estimation by using the FOBOS algorithm [Duchi and Singer, 2009b,c] and exact gradients of the log-likelihood of the Ising model. Since the computation of the exact gradient at each step is NP-hard, we used this method only for $n \leq 15$. For $n > 15$ players, we use the Höfling-Tibshirani method [Höfling and Tibshirani, 2009], which uses a sequence of first-order approximations of the exact log-likelihood. We also used a two-step algorithm, by first learning the structure by ℓ_1-regularized logistic regression [Wainwright et al., 2006] and then using the FOBOS algorithm [Duchi and Singer, 2009b,c] with belief propagation for gradient approximation. We did not find a statistically significant difference between the test log-likelihood of both algorithms and therefore we only report the latter.

Our experimental setup is as follows: after learning a model for different values of the regularization parameter ρ in a training set, we select the value of ρ that maximizes the log-likelihood in a validation set, and report statistics in a test set. For synthetic experiments, we report the Kullback-Leibler (KL) divergence, average precision (one minus the fraction of falsely included equilibria), average recall (one minus the fraction of falsely excluded equilibria) in order to measure the closeness of the recovered models to the ground truth. For real-world experiments, we report the log-likelihood. In both synthetic and real-world experiments, we report the number of equilibria and the empirical proportion of equilibria.

We first test the ability of the proposed methods to recover the ground truth structure from data. We use a small first synthetic model in order to compare with the (super-exponential) exhaustive search method. The ground truth model $\mathcal{G}_g = (\mathbf{W}_g, \mathbf{b}_g)$ has $n = 4$ players and 4 Nash equilibria (i.e. $\pi(\mathcal{G}_g)$=0.25), $\mathbf{W}_g$ was set according to Figure 7.3 (the weight of each edge was set to $+1$) and $\mathbf{b}_g = \mathbf{0}$. The mixture parameter of the ground truth q_g was set to 0.5,0.7,0.9. For each of 50 repetitions, we generated a training, a validation and a test set of 50 samples each. Figure 7.3 shows that our convex loss methods and sigmoidal maximum likelihood outperform (lower KL) exhaustive search, sigmoidal maximum empirical proportion of equilibria and Ising models. Note that the exhaustive search method which performs exact maximum likelihood suffers from over-fitting and consequently does not produce the lowest KL. From all convex loss methods, simultaneous logistic regression achieves the lowest KL. For all methods, the recovery of equilibria is perfect for $q_g = 0.9$ (number of equilibria equal to the ground truth, equilibrium precision and recall equal to 1). Additionally, the empirical proportion of equilibria resembles the mixture parameter of the ground truth q_g.

Next, we use a relatively larger second synthetic model with more complex interactions. We still keep the model small enough in order to compare with the (NP-hard) sigmoidal maximum likelihood method. The ground truth model $\mathcal{G}_g = (\mathbf{W}_g, \mathbf{b}_g)$ has $n = 9$ players and 16 Nash equilibria (i.e. $\pi(\mathcal{G}_g)$=0.03125), $\mathbf{W}_g$ was set according to Figure 7.4 (the weight of each blue and red edge was set to $+1$ and -1 respectively) and $\mathbf{b}_g = \mathbf{0}$. The mixture parameter of the ground truth q_g was set to 0.5,0.7,0.9. For each of 50 repetitions, we generated a training, a validation and a test set of 50 samples each. Figure 7.4 shows that our convex loss methods outperform (lower KL) sigmoidal methods and Ising models. From all convex loss methods, simultaneous logistic regression achieves the lowest KL. For convex loss methods, the equilibrium recovery is better than the remaining methods (number of equilibria equal to the ground truth, higher equilibrium precision and recall). Additionally, the empirical proportion of equilibria resembles the mixture parameter of the ground truth q_g.

In the next experiment, we show that the performance of convex loss minimization improves as the number of samples increases. We used random graphs with slightly more variables and varying number of samples (10,30,100,300). The ground truth model $\mathcal{G}_g = (\mathbf{W}_g, \mathbf{b}_g)$ contains $n = 20$ players. For each of 20 repetitions, we generate edges in the ground truth model $\mathbf{W}_g$ with a required density (either 0.2,0.5,0.8). For simplicity, the weight of each edge is set to $+1$ with probability $P(+1)$ and to -1 with probability $1 - P(+1)$. Hence, the Nash equilibria of the generated games does not depend on the magnitude of the weights, just on their sign. We set the bias $\mathbf{b}_g = \mathbf{0}$ and the mixture parameter of the ground truth $q_g = 0.7$. We then generated a training and a validation set with the same number of samples. Figure 7.5 shows that our convex loss methods outperform (lower KL) sigmoidal maximum empirical proportion of equilibria and Ising models (except for the synthetic model with high true proportion of equilibria: density 0.8, $P(+1) = 0$, NE$>$ 1000). The results are remarkably better when the number of equilibria in the ground truth model is small (e.g. for NE$<$ 20). From all convex loss methods, simultaneous logistic regression achieves the lowest KL.

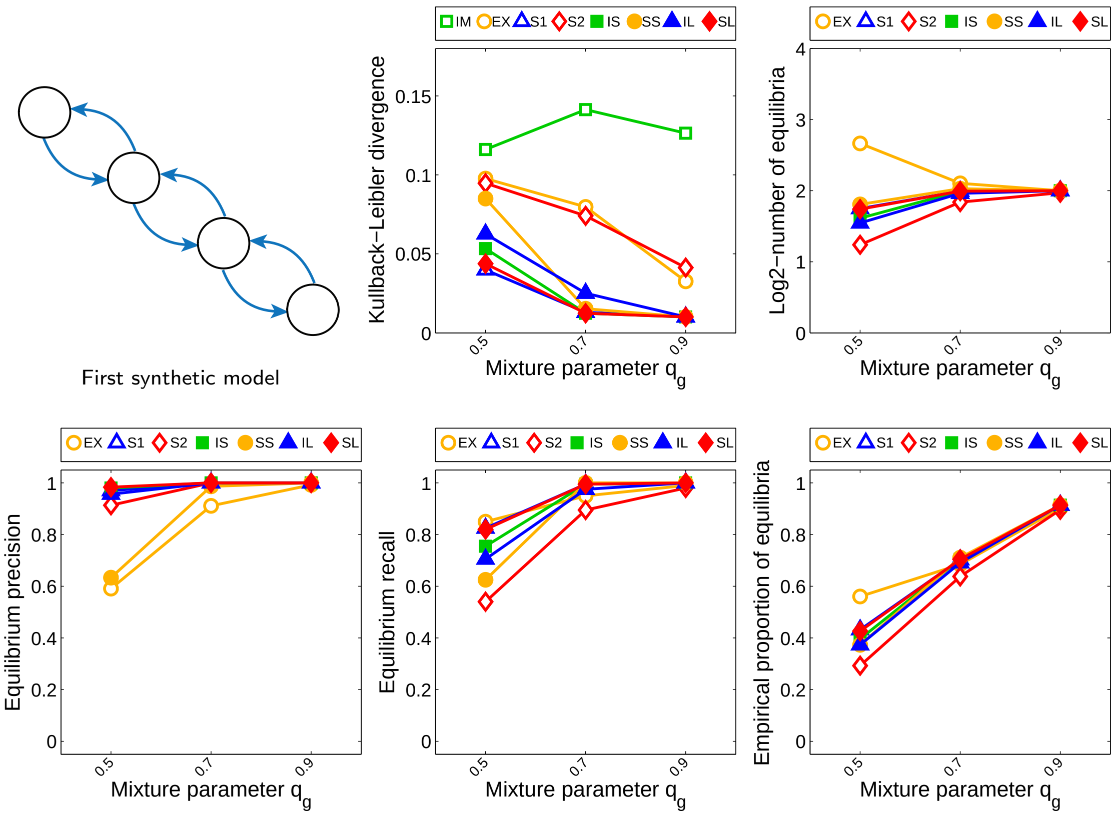


Figure 7.3: Closeness of the recovered models to the ground truth synthetic model for different mixture parameters q_g. Our convex loss methods (IS,SS: independent and simultaneous SVM, IL,SL: independent and simultaneous logistic regression) and sigmoidal maximum likelihood (S1) have lower KL than exhaustive search (EX), sigmoidal maximum empirical proportion of equilibria (S2) and Ising models (IM). For all methods, the recovery of equilibria is perfect for $q_g = 0.9$ (number of equilibria equal to the ground truth, equilibrium precision and recall equal to 1) and the empirical proportion of equilibria resembles the mixture parameter of the ground truth q_g.

In the next experiment, we evaluate two effects in our approximation methods. First, we evaluate the impact of removing the true proportion of equilibria from our objective function, i.e. the use of maximum empirical proportion of equilibria instead of maximum likelihood. Second, we evaluate the impact of using convex losses instead of a sigmoidal approximation of the 0/1 loss. We used random graphs with varying number of players and 50 samples. The ground truth model $\mathcal{G}_g = (\mathbf{W}_g, \mathbf{b}_g)$ contains $n = 4, 6, 8, 10, 12$ players. For each of 20 repetitions, we generate edges in the ground truth model $\mathbf{W}_g$ with a required density (either 0.2,0.5,0.8). As in the previous experiment, the weight of each edge is set to $+1$ with probability $P(+1)$ and to -1 with probability $1 - P(+1)$. We set the bias $\mathbf{b}_g = \mathbf{0}$ and the mixture parameter of the ground truth $q_g = 0.7$. We then generated a training and a validation set with the same number of samples. Figure 7.6 shows that in general, convex loss methods outperform (lower KL) sigmoidal maximum empirical proportion of equilibria, and the latter one outperforms sigmoidal maximum likelihood. A different effect is observed for mild (0.5) to high (0.8) density and $P(+1) = 1$ in which the sigmoidal maximum likelihood obtains the lowest KL. In a closer inspection, we found that the ground truth games usually

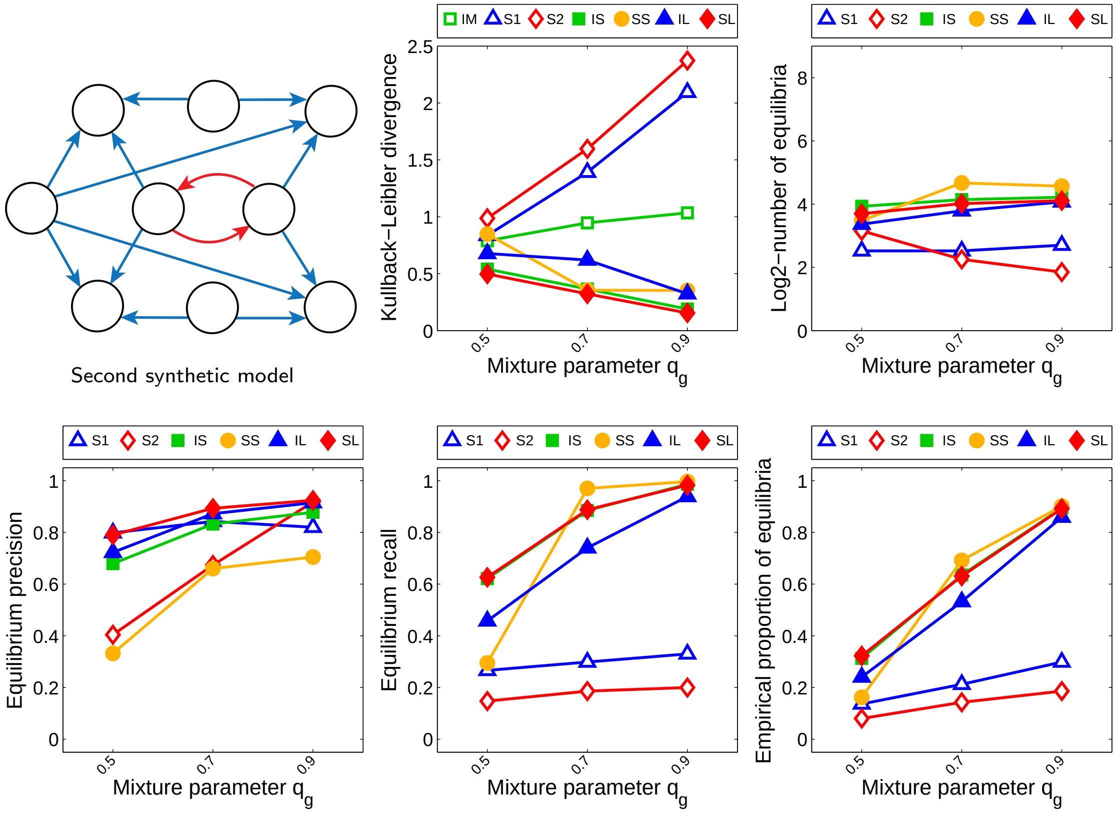


Figure 7.4: Closeness of the recovered models to the ground truth synthetic model for different mixture parameters q_g. Our convex loss methods (IS,SS: independent and simultaneous SVM, IL,SL: independent and simultaneous logistic regression) have lower KL than sigmoidal maximum likelihood (S1), sigmoidal maximum empirical proportion of equilibria (S2) and Ising models (IM). For convex loss methods, the equilibrium recovery is better than the remaining methods (number of equilibria equal to the ground truth, higher equilibrium precision and recall) and the empirical proportion of equilibria resembles the mixture parameter of the ground truth q_g.

have only 2 equilibria: $(+1, \ldots, +1)$ and $(-1, \ldots, -1)$, which seems to present a challenge for convex loss methods. It seems that for these specific cases, removing the true proportion of equilibria from the objective function negatively impacts the estimation process, but note that sigmoidal maximum likelihood is not computationally feasible for $n > 15$.

We used the U.S. congressional voting records in order to measure the generalization performance of convex loss minimization in a real-world dataset. The dataset is publicly available at `http://www.senate.gov/`. We used the first session of the 104th congress (Jan 1995 to Jan 1996, 613 votes), the first session of the 107th congress (Jan 2001 to Dec 2001, 380 votes) and the second session of the 110th congress (Jan 2008 to Jan 2009, 215 votes). Following on other researchers who have experimented with this data set (e.g. Banerjee et al. [2008]), abstentions were replaced with negative votes. Since reporting the log-likelihood requires computing the number of equilibria (which is NP-hard), we selected only 20 senators by stratified random sampling. We randomly split the data into three parts. We performed six repetitions by making each third of the data take turns as training, validation and testing sets. Figure 7.7 shows that our convex loss methods outperform (higher log-likelihood)

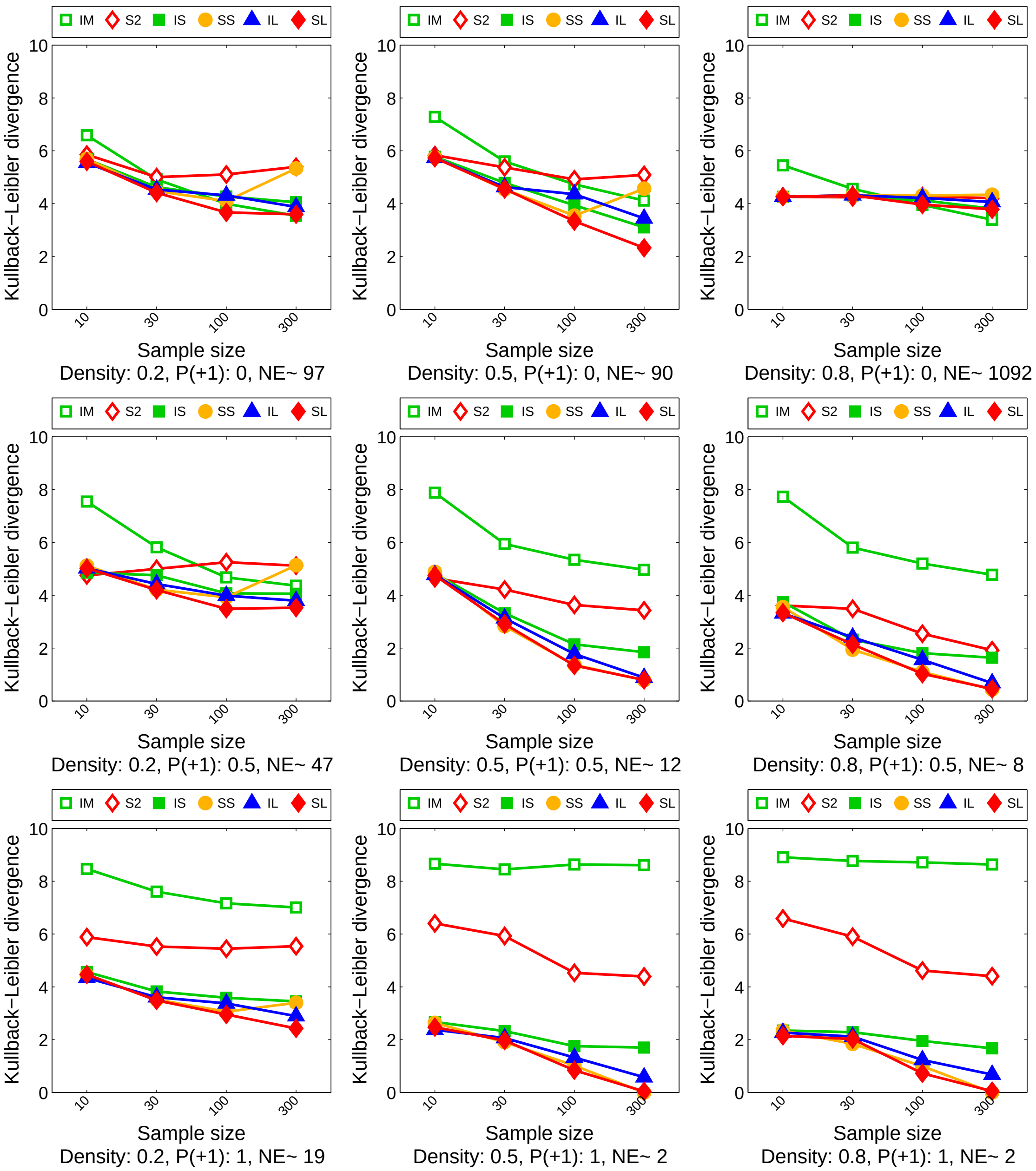
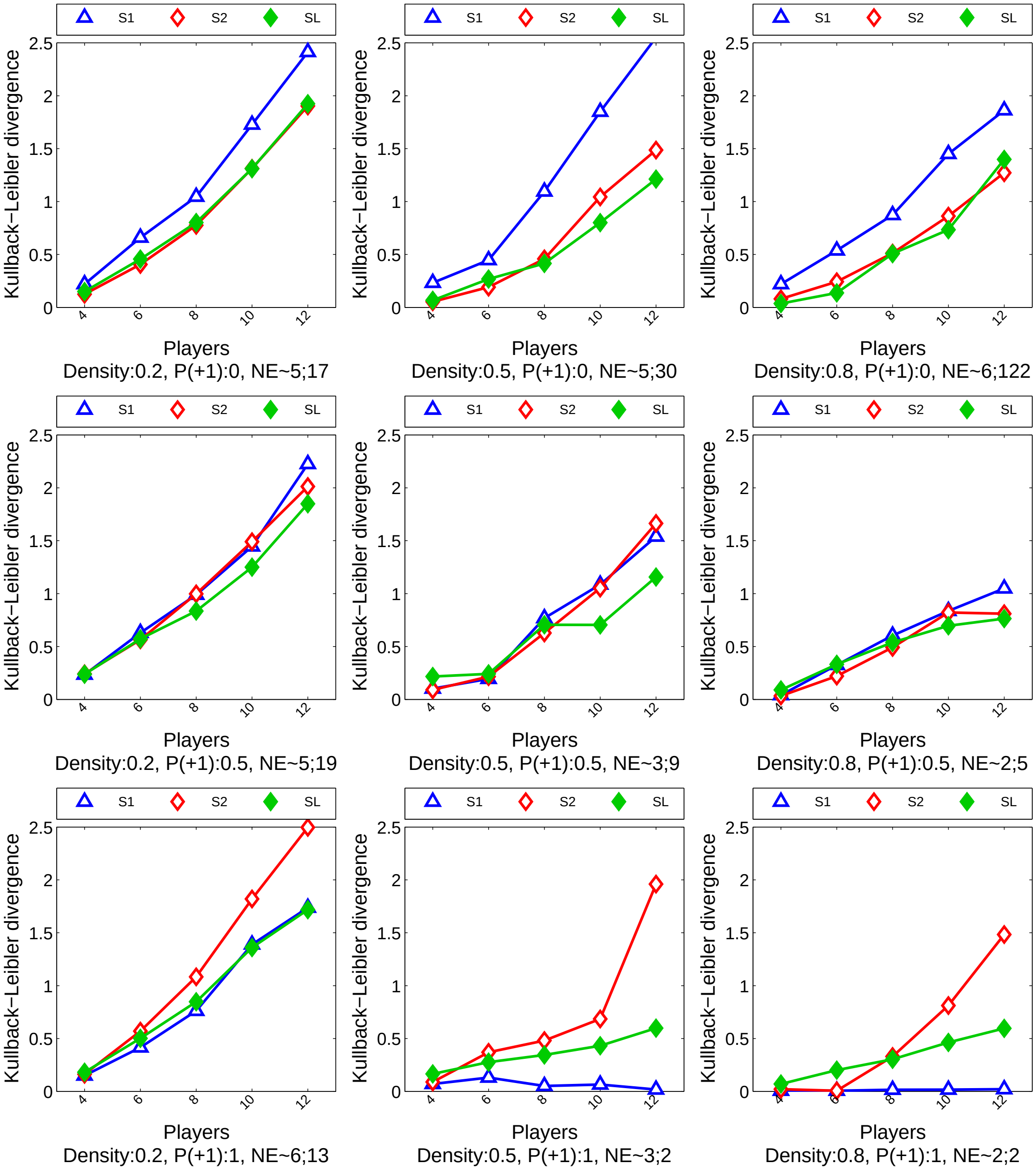


Figure 7.5: KL divergence between the recovered models and the ground truth for datasets of different number of samples. Each chart shows the density of the ground truth, probability $P(+1)$ that an edge has weight $+1$, and average number of equilibria (NE). Our convex loss methods (IS,SS: independent and simultaneous SVM, IL,SL: independent and simultaneous logistic regression) have lower KL than sigmoidal maximum empirical proportion of equilibria (S2) and Ising models (IM). The results are remarkably better when the number of equilibria in the ground truth model is small (e.g. for NE<20).

sigmoidal maximum empirical proportion of equilibria and Ising models. From all convex loss methods, simultaneous logistic regression achieves the lowest KL. For all methods, the number of equilibria (and so the true proportion of equilibria) is low.

We apply convex loss minimization to larger problems, by learning structures of games



Density:0.2, P(+1):0, NE~5;17

Density:0.5, P(+1):0, NE~5;30

Density:0.8, P(+1):0, NE~6;122

Density:0.2, P(+1):0.5, NE~5;19

Density:0.5, P(+1):0.5, NE~3;9

Density:0.8, P(+1):0.5, NE~2;5

Density:0.2, P(+1):1, NE~6;13

Density:0.5, P(+1):1, NE~3;2

Density:0.8, P(+1):1, NE~2;2

Figure 7.6: KL divergence between the recovered models and the ground truth for datasets of different number of players. Each chart shows the density of the ground truth, probability $P(+1)$ that an edge has weight $+1$, and average number of equilibria (NE) for $n = 2$;$n = 14$. In general, simultaneous logistic regression (SL) has lower KL than sigmoidal maximum empirical proportion of equilibria (S2), and the latter one has lower KL than sigmoidal maximum likelihood (S1). Other convex losses behave the same as simultaneous logistic regression (omitted for clarity of presentation).

from all 100 senators. Figure 7.8 shows that simultaneous logistic regression produce structures that are sparser than its independent counterpart. The simultaneous method better elicits the bipartisan structure of the congress. We define the influence of player j to all other players as $\sum_i |w_{ij}|$ after normalizing all weights, i.e. for each player i we divide $(\mathbf{w}_{i,-i}, b_i)$ by

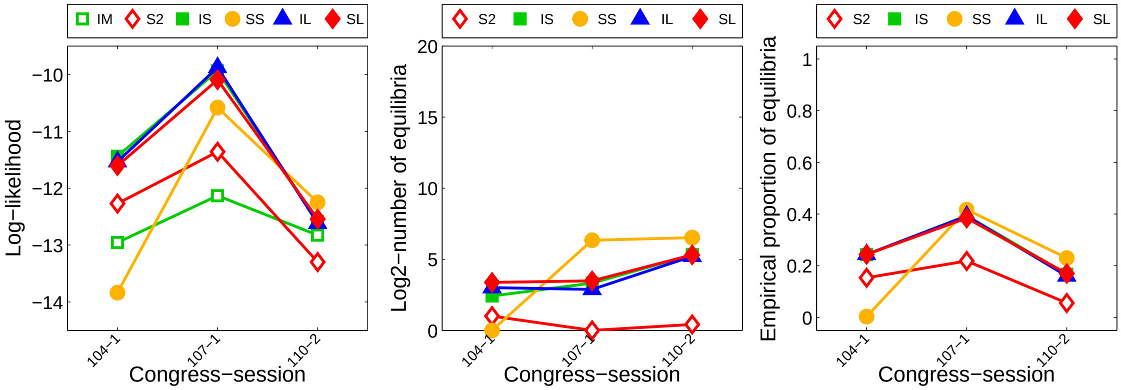


Figure 7.7: Statistics for games learnt from 20 senators from the first session of the 104th congress, first session of the 107th congress and second session of the 110th congress. The log-likelihood of our convex loss methods (IS,SS: independent and simultaneous SVM, IL,SL: independent and simultaneous logistic regression) is higher than sigmoidal maximum empirical proportion of equilibria (S2) and Ising models (IM). For all methods, the number of equilibria (and so the true proportion of equilibria) is low.

$\|\mathbf{w}_{i,-i}\|_1 + |b_i|$. Note that Jeffords and Clinton are one of the 5 most directly-influential as well as 5 least directly-influenceable (high bias) senators, in the 107th and 110th congress respectively. McCain and Feingold are both in the list of 5 most directly-influential senators in the 104th and 107th congress. McCain appears again in the list of 5 least influenciable senators in the 110th congress.

We test the hypothesis that influence between senators of the same party are stronger than senators of different party. We learn structures of games from all 100 senators from the 101th congress to the 111th congress (Jan 1989 to Dec 2010). The number of votes casted for each session were average: 337, minimum: 215, maximum: 613. Figure 7.9 validates our hypothesis and more interestingly, it shows that influence between different parties is decreasing over time. Note that the influence from Obama to Republicans increased in the last sessions, while McCain's influence to Republicans decreased.

7.9 Discussion

It is important to point out that our work is not in competition with the work in probabilistic graphical models, e.g. Ising models. Our goal is to learn the structure and parameters of games from data, and for this end, we propose a probabilistic model that is inspired by the concept of equilibrium in game theory. While we illustrate the benefit of our model in the U.S. congressional voting records, we believe that each model has its own benefits. If the practitioner "believes" that the data at hand is generated by a class of models, then the interpretation of the learnt model allows obtaining insight of the problem at hand. Note that none of the existing models (including ours) can be validated as the ground truth model that generated the real-world data, or as being more or less "realistic" with respect to other model. While generalization in unseen data is a very important measurement, a model with better generalization is not the "ground truth model" of the real-world data at hand.

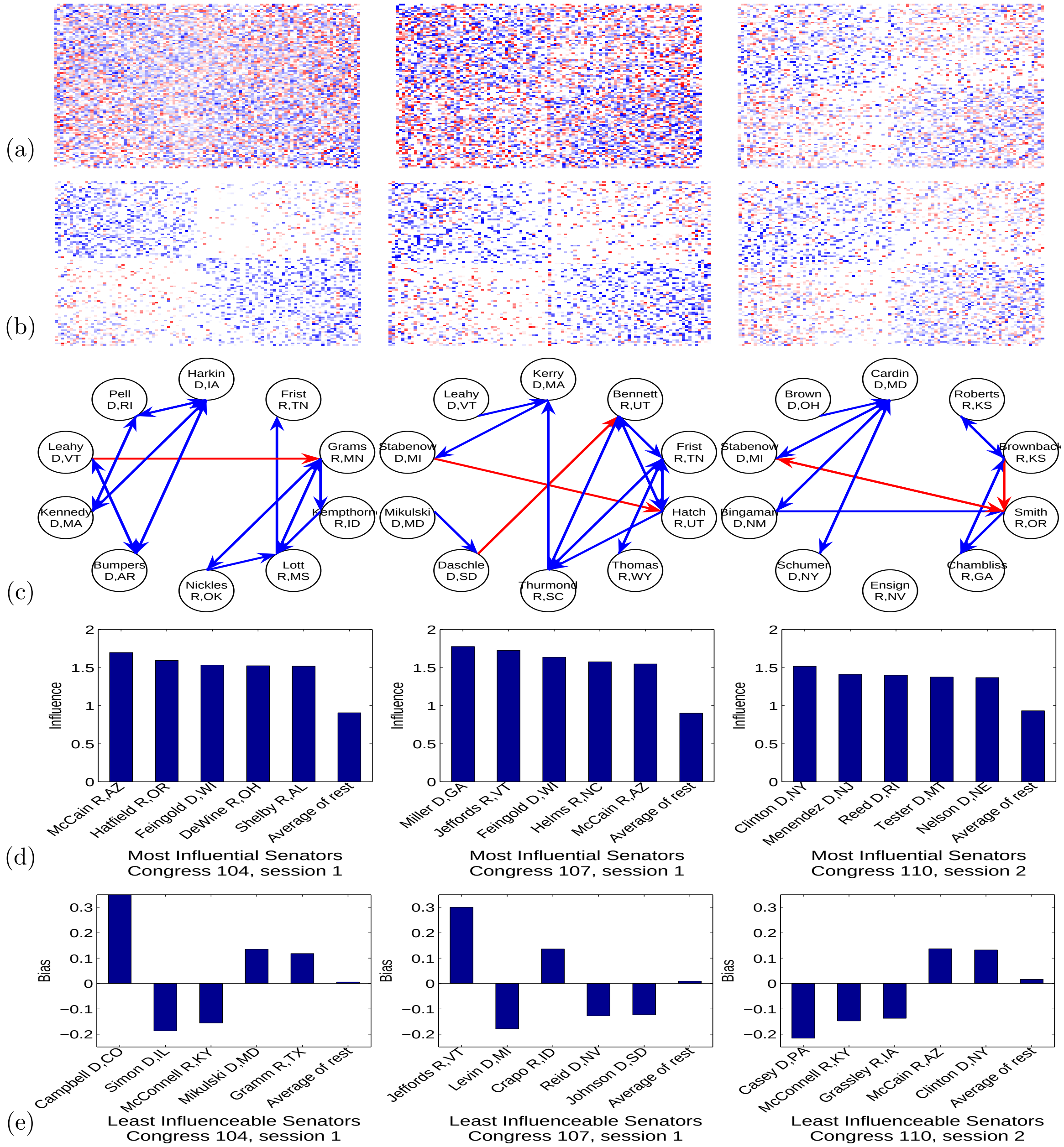


Figure 7.8: Matrix of influence weights for games learnt from all 100 senators, from the first session of the 104th congress (left), first session of the 107th congress (center) and second session of the 110th congress (right), by using our independent (a) and simultaneous (b) logistic regression methods. A row represents how every other senator influence the senator in such row. Positive influences are shown in blue, negative influences are shown in red. Democrats are shown in the top/left corner, while Republicans are shown in the bottom/right corner. Note that simultaneous method produce structures that are sparser than its independent counterpart. Partial view of the graph for simultaneous logistic regression (c). Most directly-influential (d) and least directly-influenceable (e) senators. Regularization parameter $\rho = 0.0006$.

Finally, while our model is simple, it is well founded and we show that it is far from being computationally trivial. Therefore, we believe it has its own right to be analyzed.

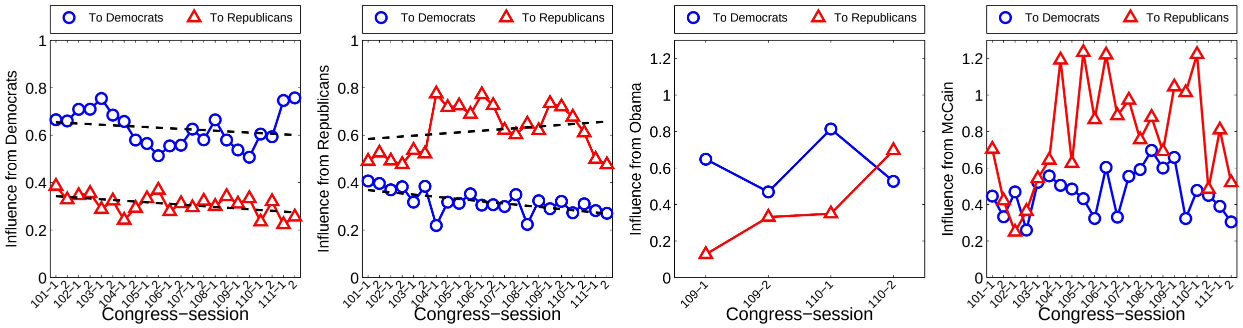


Figure 7.9: Direct influence between parties and influences from Obama and McCain. Games were learnt from all 100 senators from the 101th congress (Jan 1989) to the 111th congress (Dec 2010) by using our simultaneous logistic regression method. Direct influence between senators of the same party are stronger than senators of different party, which is also decreasing over time. In the last sessions, influence from Obama to Republicans increased, and influence from McCain to both parties decreased. Regularization parameter $\rho = 0.0006$.

The special class of graphical games considered here is related to the well-known *linear threshold model (LTM)* in sociology [Granovetter, 1978], recently very popular within the social network and theoretical computer science community [Kleinberg, 2007]. LTMs are usually studied as the basis for some kind of diffusion process. A typical problem is the identification of most influential individuals in a social network. An LTM is not in itself a game-theoretic model and, in fact, Granovetter himself argues against this view in the context of the setting and the type of questions in which he was most interested [Granovetter, 1978]. To the best of our knowledge, subsequent work on LTMs has not taken a strictly game-theoretic view either. Our model is also related to a particular model of *discrete choice with social interactions* in econometrics (see, e.g. Brock and Durlauf [2001]). The main difference is that we take a strictly non-cooperative game-theoretic approach within the classical "static"/one-shot game framework and do not use a *random utility model*. In addition, we do not make the assumption of *rational expectations*, which is equivalent to assuming that all players use exactly the same mixed strategy. As an aside note, regarding learning of information diffusion models over social networks, [Saito et al., 2010] considers a dynamic (continuous time) LTM that has only positive influence weights and a randomly generated threshold value.

There is still quite a bit of debate as to the appropriateness of game-theoretic equilibrium concepts to model individual human behavior in a social context. Camerer's book on behavioral game theory [Camerer, 2003] addresses some of the issues. We point out that there is a broader view of behavioral data, beyond those generated by individual human behavior (e.g. institutions such as nations and industries, or engineered systems such as autonomous-response devices in residential or commercial properties that are programmed to control electricity usage based on user preferences). Our interpretation of Camerer's position is not that Nash equilibiria is universally a bad predictor but that it is not *consistently* the best, for reasons that are still not well understood. This point is best illustrated in Chapter 3, Figure 3.1 of Camerer [2003]. *Quantal response equilibria (QRE)* has been proposed as an alternative to Nash in the context of behavioral game theory. Models based on QRE have been shown superior during *initial play* in some experimental settings, but most experimental work assume that the game's payoff matrices are *known* and only the "precision parameter" is estimated, e.g. Wright and Leyton-Brown [2010]. Finally, most of the human-subject experiments in behavioral game theory involve only a handful of players, and the scalability of those results to games with *many* players is unclear.

In this work we considered pure-strategy Nash equilibria only. Note that the universality of mixed-strategy Nash equilibria does not diminish the importance of pure-strategy equilibria in game theory. Indeed, a debate still exist within the game theory community as to the justification for randomization, specially in human contexts. We decided to ignore mixed-strategies due to the significant added complexity. Note that we learn exclusively from observed joint-actions, and therefore we cannot assume knowledge of the internal mixed-strategies of players. We could generalize our model to allow for mixed-strategies by defining a process in which a joint mixed strategy $\mathcal{P}$ from the set of mixed-strategy Nash equilibrium (or its complement) is drawn according to some distribution, then a (pure-strategy) realization $\mathbf{x}$ is drawn from $\mathcal{P}$ that would correspond to the observed joint-actions.

In this chapter we considered a "global" noise process, which is governed with a probability q of selecting an equilibrium. Potentially better and more natural "local" noise processes are possible, at the expense of producing a significantly more complex generative model than the one considered in this chapter. For instance, we could use a noise process that is formed of many independent, individual noise processes, one for each player. As an example, consider a the generative model in which we first select an equilibrium $\mathbf{x}$ of the game and then each player i, independently, acts according to x_i with probability q_i and switches its action with probability $1 - q_i$. The problem with such a model is that it leads to a significantly more complex expression for the generative model and thus likelihood functions. This is in contrast to the simplicity afforded us by the generative model with a more global noise process defined above.

7.10 Concluding Remarks

There are several ways of extending this research. We can extend our approach to ϵ-approximate pure-strategy Nash equilibria. In this case, for each player instead of one condition, we will have two best-response conditions which are still linear in $\mathbf{W}$ and $\mathbf{b}$. Additionally, we can extend our approach to a broader class of graphical games and non-Boolean actions. Note that our analysis does not rely on binary actions, but on binary features of one player $1[x_i = 1]$ or two players $1[x_i = x_j]$. We can use features of three players $1[x_i = x_j = x_k]$ or of non-Boolean actions $1[x_i = 3, x_j = 7]$. This kernelized version is still linear $\mathbf{W}$ and $\mathbf{b}$. These extensions are possible since our algorithms and analysis rely on linearity and binary features, additionally the VC-dimension can be modified by changing the inputs of the neural networks.

More sophisticated noise processes as well as mixed-strategy Nash equilibria need to be considered and studied. Different upper bounds for the 0/1 loss (e.g. exponential, smooth hinge) need to be analyzed. Our approach can be easily extended to parameter learning for fixed structures by using a ℓ_2^2 regularizer instead. Finally, topic-specific and time-varying versions of our model would elicit differences in preferences and trends.

Chapter 8

Conclusions and Future Work

8.1 Learning Gaussian MRFs

We presented three regularizers for maximum likelihood estimation of Gaussian MRFs: local constancy for datasets where variables correspond to a measurement in a manifold (silhouettes, motion trajectories, 2D and 3D images) in Chapter 2; variable selection for finding few interacting nodes from datasets with thousands of variables in Chapter 3; and multi-task learning for a more efficient use of data which is available for multiple related tasks in Chapter 4. For these regularizers, we showed bounds of the eigenvalues of the optimal solution, convergence of block coordinate descent optimization, and connections to the continuous quadratic knapsack problem and the quadratic trust-region problem. We presented experimental results on a wide range of complex real-world datasets with a diverse nature of probabilistic relationships: walking video sequences, motion capture, cardiac MRI, brain fMRI, gene expression, stock prices, world weather.

There are several ways of extending this research. Regarding our local constancy prior, although the positive definiteness properties of the precision matrix as well as the optimization algorithm still hold when including operators such as the Laplacian for encouraging smoothness, benefits of such a regularization approach need to be analyzed.

In practice, our techniques converge in a small number of iterations, but a more precise analysis of the rate of convergence needs to be performed. We could generalize our results on different priors for any non-negative convex regularizer. In this general setting, we could analyze the convergence rates for block coordinate descent optimization for learning Gaussian MRFs, similar to the work on the cyclic coordinate descent method for general objectives with an ℓ_1 regularizer [Saha and Tewari, 2010]. Regarding bounds of the eigenvalues of the optimal solution, we can use the concept of conjugate functions for understanding the relationship between the primal and dual problems.

Additionally, we can analyze conditions for which the recovered edges and parameters approximate the ground truth, similar to the work on edge recovery of Ravikumar et al. [2008] and consistency of the Frobenius norm of Rothman et al. [2008]. Note that given the results of Chapter 6, consistency of the Frobenius norm implies a good generalization performance (expected log-likelihood).

Regarding our multi-task prior, we experimentally found that diagonal penalization does not lead to a better generalization performance, when compared to not penalizing the diagonals. On the other hand, our methods with and without diagonal penalization recover the ground truth edges similarly well. Note that the consistency analysis of edge recovery in [Ravikumar et al., 2008] considers the *single-task* problem without diagonal penalization. It would be interesting to theoretically analyze whether diagonal penalization hurts either edge recovery, consistency of the Frobenius norm or generalization performance. Additionally, we hope the connection to the quadratic knapsack and trust-region problems will be useful for other multi-task problems, e.g. regression.

8.2 Learning Discrete MRFs

In Chapter 5, we focused on learning sparse discrete MRFs through maximum likelihood estimation. In this case, computing the objective function as well as its gradient is NP-hard. We studied the convergence rate of stochastic optimization of exact NP-hard objectives, for which only biased estimates of the gradient are available. We provided a convergence-rate analysis of deterministic errors and extend our analysis to biased stochastic errors.

There are several ways of extending this research. Although we focused on Ising models, the ideas developed in Chapter 5 could be applied to Markov random fields with higher order cliques. Our analysis can be easily extended to parameter learning for fixed structures by using a ℓ_2^2 regularizer instead.

Although we show that accelerated proximal gradient is not guaranteed to converge in our specific biased stochastic setting, necessary conditions for its convergence needs to be investigated.

Note that our analysis used very little knowledge regarding the specifics of the (B, V, S, D)-sampler, e.g. MCMC. In fact, the results of [Peskun, 1973] apply to estimates where all the samples are generated by a single MCMC simulation, and therefore the samples are not independent. On the other hand, practitioners perform several MCMC simulations and use only the last sample from each of the simulations. It is clear that the results of [Peskun, 1973] for the *dependent samples* approach provides a worst case bound for the *independent samples* approach. It is probable that tighter results can be obtained by revisiting this problem. Besides this, we believe that by using more knowledge of the MCMC sampler (e.g. burn-in time, dependence between iterations in forward-backward splitting) we could obtain tighter and more informative convergence rates.

As we mentioned in Chapter 5, analytical approximations of the bias and variance constants (B and V) seem to be difficult to obtain even for specific classes, e.g. Ising models. More work needs to be performed along this line.

8.3 General Ideas for Learning Gaussian or Discrete MRFs

In a real-world scenario, practitioners usually mix different priors in the learning process. Block coordinate descent methods, forward-backward splitting and projected gradient methods rely on a closed-form step at each iteration. Typically, this closed-form step would need to be specifically derived for each particular mixture of priors. We wonder whether it is possible to produce a more general algorithm that guarantees convergence and, at the same time, does not require these analytical derivations. A reasonable approach would be to sequentially apply the closed-form step for the different priors in the mixture, although this seems to be a non-trivial problem since the order in which the priors are applied might affect the final solution.

Another very interesting line of research would be to produce algorithms for learning graphical models with continuous and discrete variables.

For real-world data, we used the test log-likelihood in order to measure the generalization performance and the test classification accuracy to measure the discriminability of the models. For synthetic data, we used ROC curves to measure the quality of edge recovery. From a methodological point of view, we wonder whether it is possible to devise a method to measure the quality of edge recovery in real-world data.

8.4 Lipschitz Parameterization of Probabilistic Graphical Models

In Chapter 6, we showed general results for graphical models that allow understanding maximum likelihood estimation with regularizers on the differences of parameters, the generalization ability of graphical models, and the use of model parameters as features in classification, dimensionality reduction and clustering. To this end, we showed that the log-likelihood of several graphical models is Lipschitz continuous with respect to the parameters, and derived bounds on the Kullback-Leibler divergence, expected log-likelihood and Bayes error rate.

There are several ways of extending this research. We hope that our preliminary results will motivate work on proving other theoretical properties as well as on learning probabilistic graphical models by using optimization algorithms that rely on Lipschitz continuity of the log-likelihood as the objective function. Finally, while Lipschitz continuity defines an upper bound of the derivative, lower bounds of the derivative will allow for finding a lower bound of the Kullback-Leibler divergence as well as upper bounds for the Bayes error and the expected log-likelihood.

8.5 Learning Linear Influence Games

In Chapter 7, we formalized and studied the problem of learning the structure of graphical games from strictly behavioral data. We proposed maximum likelihood estimation of a generative model defined by the Nash equilibria of the game. We showed a generalization bound for maximum likelihood estimation. We discuss several optimization algorithms including convex loss minimization, sigmoidal approximations and exhaustive search. We formally prove that games in our hypothesis space have a small true number of equilibria, with high probability; thus, convex loss minimization is sound. Finally, we provided experimental results on the U.S. congressional voting records.

There are several ways of extending this research. We can easily extend our approach to ϵ-approximate pure-strategy Nash equilibria. In our analysis, we used interactions between two players and binary actions. We can extend our approach to a broader class of graphical games and non-Boolean actions. This broader class of graphical games can include interactions of more than two players and general discrete actions. As we argued in Chapter 7 our analysis only relies on binary "features" and linearity of the payoff function with respect to the model parameters, therefore a kernelized version of our approach is very likely to work.

More sophisticated noise processes as well as mixed-strategy Nash equilibria need to be considered and studied. Different upper bounds for the 0/1 loss (e.g. exponential, smooth hinge) need to be analyzed. Our approach can be easily extended to parameter learning for fixed structures by using a ℓ_2^2 regularizer instead. Finally, topic-specific and time-varying versions of our model would elicit differences in preferences and trends.

Bibliography

P. Abbeel, D. Koller, and A. Ng. Learning Factor Graphs in Polynomial Time and Sample Complexity. *UAI*, 2005.

O. Aichholzer and F. Aurenhammer. Classifying Hyperplanes in Hypercubes. *SIAM Journal on Discrete Mathematics*, 9(2):225–232, 1996.

A. Asuncion, Q. Liu, A. Ihler, and P. Smyth. Particle Filtered MCMC-MLE with Connections to Contrastive Divergence. *ICML*, 2010.

R. Aumann. Subjectivity and Correlation in Randomized Strategies. *Journal of Mathematical Economics*, 1, 1974.

M. Baes. Estimate sequence methods: extensions and approximations. *IFOR internal report, ETH Zurich*, 2009.

O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data. *JMLR*, 2008.

O. Banerjee, L. El Ghaoui, A. d'Aspremont, and G. Natsoulis. Convex Optimization Techniques for Fitting Sparse Gaussian Graphical Models. *ICML*, 2006.

F. Barahona. On the computational complexity of Ising spin glass models. *Journal of Physics A: Mathematical, Nuclear and General*, 1982.

J. Besag. Statistical Analysis of Non-Lattice Data. *The Statistician*, 1975.

C. Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006.

B. Blum, C.R. Shelton, and D. Koller. A Continuation Method for Nash Equilibria in Structured Games. *JAIR*, 25:457–502, 2006.

S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2006.

P. Bradley and O. Mangasarian. Feature Selection Via Concave Minimization and Support Vector Machines. *ICML*, 1998.

W. Brock and S. Durlauf. Discrete Choice with Social Interactions. *The Review of Economic Studies*, 68(2):235–260, 2001.

F. Bromberg, D. Margaritis, and V. Honavar. Efficient Markov Network Structure Discovery Using Independence Tests. *SIAM International Conference on Data Mining*, 2006.

P. Brucker. An $O(n)$ algorithm for quadratic knapsack problems. *Operations Research Letters*, 1984.

C. Camerer. *Behavioral Game Theory: Experiments on Strategic Interaction*. Princeton University Press, 2003.

G. Cecchi, I. Rish, B. Thyreau, B. Thirion, M. Plaze, M. Paillere-Martinot, C. Martelli, J. Martinot, and J. Poline. Discriminative Network Models of Schizophrenia. *NIPS*, 2009.

A. Chan, N. Vasconcelos, and G. Lanckriet. Direct Convex Relaxations of Sparse SVM. *ICML*, 2007.

V. Chandrasekaran, N. Srebro, and P. Harsha. Complexity of Inference in Graphical Models. *UAI*, 2008.

R. Chellappa and S. Chatterjee. Classification of Textures Using Gaussian Markov Random Fields. *IEEE Trans. Acoustics, Speech and Signal Processing*, 1985.

C. Chen and R. Dubes. Discrete MRF model parameters as features for texture classification. *IEEE Conf. Systems, Man and Cybernetics*, 1990.

X. Chen, Q. Lin, S. Kim, J. Carbonell, and E. Xing. Smoothing Proximal Gradient Method for General Structured Sparse Learning. *UAI*, 2011.

H. Chernoff. A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the sum of Observations. *The Annals of Mathematical Statistics*, 1952.

C. Chow and C. Liu. Approximating Discrete Probability Distributions with Dependence Trees. *IEEE Transactions on Information Theory*, 1968.

D. Crandall, P. Felzenszwalb, and D. Huttenlocher. Spatial priors for part-based recognition using statistical models. *CVPR*, 2005.

C. Daskalakis, A. Dimakisy, and E. Mossel. Connectivity and Equilibrium in Random Games. *Annals of Applied Probability*, 21(3):987–1016, 2011.

C. Daskalakis, P. Goldberg, and C. Papadimitriou. The complexity of computing a Nash equilibrium. *Commun. ACM*, 52(2):89–97, 2009.

A. d'Aspremont. Smooth optimization with approximate gradient. *SIAM Journal on Optimization*, 2008.

A. Dempster. Covariance Selection. *Biometrics*, 1972.

G. Desjardins, A. Courville, Y. Bengio, P. Vincent, and O. Delalleau. Parallel Tempering for Training of Restricted Boltzmann Machines. *AISTATS*, 2010.

J. Deux, A. Rahmouni, and J. Garot. Cardiac magnetic resonance and 64-slice cardiac CT of lipomatous metaplasia of chronic myocardial infarction. *European Heart Journal*, 2008.

O. Devolder. Stochastic First Order Methods in Smooth Convex Optimization. *CORE Discussion Papers 2012/9*, 2012.

O. Devolder, F. Glineur, and Y. Nesterov. First-order methods of smooth convex optimization with inexact oracle. *CORE Discussion Papers 2011/2*, 2011.

P. Domingos and M. Pazzani. On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning*, 1997.

J. Duchi, A. Agarwal, M. Johansson, and M. Jordan. Ergodic subgradient descent. *Allerton Conference*, 2011.

J. Duchi, S. Gould, and D. Koller. Projected Subgradient Methods for Learning Sparse Gaussians. *UAI*, 2008a.

J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient Projections onto the ℓ_1-Ball for Learning in High Dimensions. *ICML*, 2008b.

J. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite Objective Mirror Descent. *COLT*, 2010.

J. Duchi and Y. Singer. Boosting with Structural Sparsity. *ICML*, 2009a.

J. Duchi and Y. Singer. Efficient Learning using Forward-Backward Splitting. *NIPS*, 2009b.

J. Duchi and Y. Singer. Efficient Online and Batch Learning using Forward Backward Splitting. *JMLR*, 2009c.

R. Duda, P. Hart, and D. Stork. *Pattern Classification.* Wiley-Interscience, 2001.

Q. Duong, Y. Vorobeychik, S. Singh, and M. Wellman. Learning Graphical Game Models. *IJCAI*, 2009.

L. El Ghaoui and A. Gueye. A Convex Upper Bound on the Log-Partition Function for Binary Graphical Models. *NIPS*, 2008.

G. Elidan and N. Friedman. Learning Hidden Variable Networks: The Information Bottleneck Approach. *JMLR*, 2005.

A. Fabrikant, C. Papadimitriou, and K. Talwar. The complexity of pure Nash equilibria. *STOC*, 2004.

P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 2005.

S. Ficici, D. Parkes, and A. Pfeffer. Learning and Solving Many-Player Games through a Cluster-Based Representation. *UAI*, 2008.

G. Forsythe and G. Golub. On the Stationary Values of a Second-Degree Polynomial on the Unit Sphere. *SIAM Journal of the Society for Industrial and Applied Mathematics*, 1965.

B. Frey and N. Jojic. A Comparison of Algorithms for Inference and Learning in Probabilistic Graphical Models. *PAMI*, 2005.

M. Friedlander and M. Schmidt. Hybrid deterministic-stochastic methods for data fitting. *arXiv:1104.2373*, 2011.

J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise Coordinate Optimization. *The Annals of Applied Statistics*, 2007a.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse Inverse Covariance Estimation with the Graphical Lasso. *Biostatistics*, 2007b.

J. Friedman, T. Hastie, and R. Tibshirani. Applications of the lasso and grouped lasso to the estimation of sparse graphical models. *Technical Report, Stanford University*, 2010.

N. Friedman. Learning Belief Networks in the Presence of Missing Values and Hidden Variables. *ICML*, 1997.

N. Friedman. The Bayesian Structural EM Algorithm. *UAI*, 1998.

N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian Network Classifiers. *Machine Learning*, 1997.

D. Fudenberg and J. Tirole. *Game Theory.* The MIT Press, 1991.

X. Gao and A. Pfeffer. Learning Game Representations from Data Using Rationality Constraints. *UAI*, 2010.

C. Geyer. Markov chain Monte Carlo maximum likelihood. *Computing Science and Statistics*, 1991.

R. Goldstein, N. Alia-Klein, D. Tomasi, J. Honorio, T. Maloney, P. Woicik, R. Wang, F. Telang, and N. Volkow. Anterior cingulate cortex hypoactivations to an emotionally salient task in cocaine addiction. *Proceedings of the National Academy of Sciences, USA*, 2009.

R. Goldstein, D. Tomasi, N. Alia-Klein, L. Zhang, F. Telang, and N. Volkow. The effect of practice on a sustained attention task in cocaine abusers. *NeuroImage*, 2007.

M. Granovetter. Threshold Models of Collective Behavior. *The American Journal of Sociology*, 83(6):1420–1443, 1978.

L. Gu, E. Xing, and T. Kanade. Learning GMRF Structures for Spatial Priors. *CVPR*, 2007.

J. Guo, E. Levina, G. Michailidis, and J. Zhu. Joint Estimation of Multiple Graphical Models. *Biometrika*, 2010.

Y. Guo and D. Schuurmans. Convex Structure Learning for Bayesian Networks: Polynomial Feature Selection and Approximate Ordering. *UAI*, 2006.

K. Helgason, J. Kennington, and H. Lall. A polynomially bounded algorithm for a singly constrained quadratic program. *Mathematical Programming*, 1980.

G. Hinton. Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation*, 2002.

H. Höfling and R. Tibshirani. Estimation of Sparse Binary Pairwise Markov Networks using Pseudo-likelihoods. *JMLR*, 2009.

J. Honorio. Lipschitz Parametrization of Probabilistic Graphical Models. *UAI*, 2011.

J. Honorio. Convergence Rates of Biased Stochastic Optimization for Learning Sparse Ising Models. *ICML*, 2012.

J. Honorio and L. Ortiz. Learning the Structure of Large-Population Graphical Games from Behavioral Data. *soon to be submitted, arXiv:1206.3713*, 2012.

J. Honorio, L. Ortiz, D. Samaras, N. Paragios, and R. Goldstein. Sparse and Locally Constant Gaussian Graphical Models. *NIPS*, 2009.

J. Honorio and D. Samaras. Multi-Task Learning of Gaussian Graphical Models. *ICML*, 2010.

J. Honorio and D. Samaras. Simultaneous and Group-Sparse Multi-Task Learning of Gaussian Graphical Models. *JMLR, under submission, arXiv:1207.4255*, 2012.

J. Honorio, D. Samaras, I. Rish, and G. Cecchi. Variable Selection for Gaussian Graphical Models. *AISTATS*, 2012.

C. Hu, J. Kowk, and W. Pan. Accelerated Gradient Methods for Stochastic Optimization and Online Learning. *NIPS*, 2009.

M. Irfan and L. Ortiz. A Game-Theoretic Approach to Influence in Networks. *AAAI*, 2011.

E. Ising. Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift für Physik*, 1925.

A. Jalali, C. Johnson, and P. Ravikumar. On Learning Discrete Graphical Models Using Greedy Methods. *NIPS*, 2011.

E. Janovskaja. Equilibrium situations in multi-matrix games. *Litovskiĭ Matematicheskiĭ Sbornik*, 8:381–384, 1968.

T. Jebara. Multi-Task Feature and Kernel Selection for SVMs. *ICML*, 2004.

A. Jiang and K. Leyton-Brown. Polynomial-time Computation of Exact Correlated Equilibrium in Compact Games. *ACM Electronic Commerce Conference*, 2011.

J. Johnson, V. Chandrasekaran, and A. Willsky. Learning Markov Structure by Maximum Entropy Relaxation. *AISTATS*, 2007.

S. Kakade, M. Kearns, J. Langford, and L. Ortiz. Correlated equilibria in graphical games. *ACM Electronic Commerce Conference*, 2003.

M. Kearns, M. Littman, and S. Singh. Graphical Models for Game Theory. *UAI*, 2001.

M. Kearns and J. Wortman. Learning from Collective Behavior. *COLT*, 2008.

K. Kiwiel. On Linear-Time Algorithms for the Continuous Quadratic Knapsack Problem. *Journal of Optimization Theory and Applications*, 2007.

J. Kleinberg. Cascading Behavior in Networks: Algorithmic and Economic Issues. In Noam Nisan, Tim Roughgarden, Éva Tardos, and Vijay V. Vazirani, editors, *Algorithmic Game Theory*, chapter 24, pages 613–632. Cambridge University Press, 2007.

M. Kolar, L. Song, A. Ahmed, and E. Xing. Estimating time-varying networks. *Annals of Applied Statistics*, 2010.

M. Kolar, L. Song, and E. Xing. Sparsistent Learning of Varying-coefficient Models with Structural Changes. *NIPS*, 2009.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009.

J. Langford, L. Li, and T. Zhang. Sparse Online Learning via Truncated Gradient. *JMLR*, 2009.

S. Lauritzen. *Graphical Models*. Oxford Press, 1996.

J. Lee, F. Telang, C. Springer, and N. Volkow. Abnormal brain activation to visual stimulation in cocaine abusers. *Life Sciences*, 2003.

S. Lee, V. Ganapathi, and D. Koller. Efficient Structure Learning of Markov Networks Using ℓ_1-Regularization. *NIPS*, 2006a.

S. Lee, H. Lee, P. Abbeel, and A. Ng. Efficient ℓ_1 Regularized Logistic Regression. *AAAI*, 2006b.

E. Levina, A. Rothman, and J. Zhu. Sparse Estimation of Large Covariance Matrices via a Nested Lasso Penalty. *The Annals of Applied Statistics*, 2008.

H. Liu, M. Palatucci, and J. Zhang. Blockwise Coordinate Descent Procedures for the Multi-task Lasso, with Applications to Neural Semantic Basis Discovery. *ICML*, 2009a.

J. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, 2001.

J. Liu, S. Ji, and J. Ye. Multi-Task Feature Learning Via Efficient $\ell_{2,1}$-Norm Minimization. *UAI*, 2009b.

Q. Liu and A. Ihler. Learning Scale Free Networks by Reweighted ℓ_1 regularization. *AISTATS*, 2011.

J. Mairal, R. Jenatton, G. Obozinski, and F. Bach. Network Flow Algorithms for Structured Sparsity. *NIPS*, 2010.

V. Mansinghka, C. Kemp, J. Tenenbaum, and T. Griffiths. Structured Priors for Structure Learning. *UAI*, 2006.

B. Marlin and K.Murphy. Sparse Gaussian Graphical Models with Unknown Block Structure. *ICML*, 2009.

B. Marlin, M. Schmidt, and K. Murphy. Group Sparse Priors for Covariance Estimation. *UAI*, 2009.

B. Marlin, K. Swersky, B. Chen, and N. de Freitas. Inductive Principles for Restricted Boltzmann Machine Learning. *AISTATS*, 2010.

L. Meier, S. van de Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society*, 2008.

N. Meinshausen and P. Bühlmann. High Dimensional Graphs and Variable Selection with the Lasso. *The Annals of Statistics*, 2006.

J. Moré and D. Sorensen. Computing a Trust Region Step. *SIAM Journal on Scientific and Statistical Computing*, 1983.

S. Muroga. Lower bounds on the number of threshold functions and a maximum weight. *IEEE Transactions on Electronic Computers*, 14:136148, 1965.

S. Muroga. *Threshold Logic and Its Applications.* John Wiley & Sons, 1971.

S. Muroga and I. Toda. Lower Bound of the Number of Threshold Functions. *IEEE Transactions on Electronic Computers*, 5:805–806, 1966.

I. Murray and Z. Ghahramani. Bayesian Learning in Undirected Graphical Models: Approximate MCMC algorithms. *UAI*, 2004.

J. Myers, K. Laskey, and T. Levitt. Learning Bayesian Networks from Incomplete Data with Stochastic Search Algorithms. *UAI*, 1999.

J. Nash. Non-cooperative games. *Annals of Mathematics*, 54:286–295, 1951.

G. Natsoulis, L. El Ghaoui, G. Lanckriet, A. Tolley, F. Leroy, S. Dunlea, B. Eynon, C. Pearson, S. Tugendreich, and K. Jarnagin. Classification of a large microarray data set: algorithm comparison and analysis of drug signatures. *Genome Research*, 2005.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust Stochastic Approximation Approach to Stochastic Programming. *SIAM Journal on Optimization*, 2009.

A. Niculescu-Mizil and R. Caruana. Inductive Transfer for Bayesian Network Structure Learning. *AISTATS*, 2007.

G. Obozinski, B. Taskar, and M. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 2010.

L. Ortiz and M. Kearns. Nash Propagation for Loopy Graphical Games. *NIPS*, 2002.

C. Papadimitriou and T. Roughgarden. Computing correlated equilibria in multi-player games. *Journal of the ACM*, 55(3):1–29, 2008.

S. Parise and M. Welling. Structure Learning in Markov Random Fields. *NIPS*, 2006.

R. Parr, L. Li, G. Taylor, C. Painter-Wakefield, and M. Littman. An Analysis of Linear Models, Linear Value-Function Approximation, and Feature Selection for Reinforcement Learning. *ICML*, 2008.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, 1988.

P. Peskun. Optimum Monte Carlo Sampling Using Markov Chains. *Biometrika*, 1973.

Y. Qi, D. Liu, L. Carin, and D. Dunson. Multi-Task Compressive Sensing with Dirichlet Process Priors. *ICML*, 2008.

A. Quattoni, X. Carreras, M. Collins, and T. Darrell. An Efficient Projection for $\ell_{1,\infty}$ Regularization. *ICML*, 2009.

M. Ramoni and P. Sebastiani. Learning Bayesian Networks from Incomplete Databases. *UAI*, 1997.

P. Ravikumar, G. Raskutti, M. Wainwright, and B. Yu. Model Selection in Gaussian Graphical Models: High-Dimensional Consistency of ℓ_1-regularized MLE. *NIPS*, 2008.

A. Rothman, P. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2008.

S. Roy, T. Lane, and M. Werner-Washburne. Learning structurally consistent undirected probabilistic graphical models. *ICML*, 2009.

A. Saha and A. Tewari. On the Finite Time Convergence of Cyclic Coordinate Descent Methods. *Pre-print*, 2010.

K. Saito, M. Kimura, K. Ohara, and H. Motoda. Selecting Information Diffusion Models over Social Networks for Behavioral Analysis. *ECML*, 2010.

R. Salakhutdinov. Learning in Markov Random Fields using Tempered Transitions. *NIPS*, 2009.

R. Salakhutdinov. Learning Deep Boltzmann Machines using Adaptive MCMC. *ICML*, 2010.

M. Schmidt, G. Fung, and R. Rosales. Fast Optimization Methods for ℓ_1 Regularization: A Comparative Study and Two New Approaches. *ECML*, 2007a.

M. Schmidt, N. Le Roux, and F. Bach. Convergence Rates of Inexact Proximal-Gradient Methods for Convex Optimization. *NIPS*, 2011.

M. Schmidt and K. Murphy. Modeling Discrete Interventional Data using Directed Cyclic Graphical Models. *UAI*, 2009.

M. Schmidt, K. Murphy, G. Fung, and R. Rosales. Structure Learning in Random Fields for Heart Motion Abnormality Detection. *CVPR*, 2008.

M. Schmidt, A. Niculescu-Mizil, and K. Murphy. Learning Graphical Model Structure Using ℓ_1-Regularization Paths. *AAAI*, 2007b.

M. Schmidt, E. van den Berg, M. Friedlander, and K. Murphy. Optimizing Costly Functions with Simple Constraints: A Limited-Memory Projected Quasi-Newton Algorithm. *AISTATS*, 2009.

S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal Estimated Sub-gradient Solver for SVM. *ICML*, 2007.

S. Shalev-Shwartz and A. Tewari. Stochastic Methods for ℓ_1 Regularized Loss Minimization. *ICML*, 2009.

N. Shor. *Minimization Methods for Non-differentiable Functions*. Springer-Verlag, 1985.

E. Sontag. VC Dimension of Neural Networks. In *Neural Networks and Machine Learning*, pages 69–95. Springer, 1998.

N. Srebro. Maximum Likelihood Bounded Tree-Width Markov Networks. *UAI*, 2001.

R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, 1996.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and Smoothness via the Fused Lasso. *Journal of the Royal Statistical Society*, 2005.

T. Tieleman. Training Restricted Boltzmann Machines using Approximations to the Likelihood Gradient. *ICML*, 2008.

J. Tropp. Algorithms for simultaneous sparse approximation, Part II: convex relaxation. *Signal Processing*, 2006.

P. Tseng. Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization. *Journal of Optimization Theory and Applications*, 2001.

B. Turlach, W. Venables, and S. Wright. Simultaneous variable selection. *Technometrics*, 2005.

O. Tuzel, F. Porikli, and P. Meer. Human Detection via Classification on Riemannian Manifolds. *CVPR*, 2007.

G. Varoquaux, A. Gramfort, J. Poline, and B. Thirion. Brain Covariance Selection: Better Individual Functional Connectivity Models Using Population Prior. *NIPS*, 2010.

D. Vickrey and D. Koller. Multi-Agent Algorithms for Solving Graphical Games. *AAAI*, 2002.

Y. Vorobeychik, M. Wellman, and S. Singh. Learning Payoff Functions in Infinite Games. *IJCAI*, 2005.

M. Wainwright, P. Ravikumar, and J. Lafferty. High dimensional Graphical Model Selection Using ℓ_1-Regularized Logistic Regression. *NIPS*, 2006.

K. Waugh, B. Ziebart, and J. Bagnell. Computational Rationalization: The Inverse Equilibrium Problem. *ICML*, 2011.

A. Wilson, A. Fern, S. Ray, and P. Tadepalli. Multi-Task Reinforcement Learning: A Hierarchical Bayesian Approach. *ICML*, 2007.

R. Winder. Single state threshold logic. *Switching Circuit Theory and Logical Design*, S-134: 321–332, 1960.

J. Wright and K. Leyton-Brown. Beyond Equilibrium: Predicting Human Behavior in Normal Form Games. *AAAI*, 2010.

S. Yamija and T. Ibaraki. A lower bound of the number of threshold functions. *IEEE Transactions on Electronic Computers*, 14:926929, 1965.

E. Yang and P. Ravikumar. On the Use of Variational Inference for Learning Discrete Graphical Models. *ICML*, 2011.

L. Younes. Estimation and annealing for Gibbsian fields. *Annales de l'Institut Henri Poincaré*, 1988.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society*, 2006.

M. Yuan and Y. Lin. Model Selection and Estimation in the Gaussian Graphical Model. *Biometrika*, 2007.

B. Zhang and Y. Wang. Learning Structural Changes of Gaussian Graphical Models in Controlled Experiments. *UAI*, 2010.

J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani. 1-norm Support Vector Machines. *NIPS*, 2003.

B. Ziebart, J. Bagnell, and A. Dey. Modeling Interaction via the Principle of Maximum Causal Entropy. *ICML*, 2010.

Appendix A

ℓ_1-norm Versus Squared ℓ_2-norm for Local Constancy

In order to show why we use the ℓ_1-norm penalty encouraging local constancy, we use an example borrowed from [Tibshirani et al., 2005]. This example is a simpler problem than structure learning for Gaussian graphical models, but it is tightly related and it allows us to acquire a visual grasp on comparing the use of ℓ_1-norm versus ℓ_2-norm for local constancy.

Given $N = 9$, we want to find the most sparse and locally constant profile $\mathbf{y} \in \mathbb{R}^N$ that resembles the values $\widehat{\mathbf{y}} \in \mathbb{R}^N$ as close as possible. Let $\mathbf{D} \in \mathbb{R}^{N-1 \times N}$ be the matrix corresponding to the differential operator. Using the ℓ_1-norm or squared ℓ_2-norm penalty for encouraging local constancy, we obtain:

$$\begin{array}{c} \min_{\mathbf{y} \in \mathbb{R}^N} \left(\frac{1}{2} \|\mathbf{y} - \widehat{\mathbf{y}}\|_2^2 + \rho \|\mathbf{y}\|_1 + \tau \|\mathbf{D}\mathbf{y}\|_1 \right) \text{ for the } \ell_1 - \text{norm} \\ \text{or} \\ \min_{\mathbf{y} \in \mathbb{R}^N} \left(\frac{1}{2} \|\mathbf{y} - \widehat{\mathbf{y}}\|_2^2 + \rho \|\mathbf{y}\|_1 + \tau \|\mathbf{D}\mathbf{y}\|_2^2 \right) \text{ for the } \ell_2 - \text{norm} \end{array}$$

for some $\rho, \tau > 0$. The first term $\|\mathbf{y} - \widehat{\mathbf{y}}\|_2^2$ models the quality of the fit, the second term $\rho\|\mathbf{y}\|_1$ encourages sparseness while the third term $\tau\|\mathbf{D}\mathbf{y}\|_1$ or $\tau\|\mathbf{D}\mathbf{y}\|_2^2$ encourages local constancy.

Figure A.1 shows that ℓ_1-norm penalties for local constancy lead to locally constant models which preserve sparseness, where as squared ℓ_2-norm penalties of differences fails to do so.

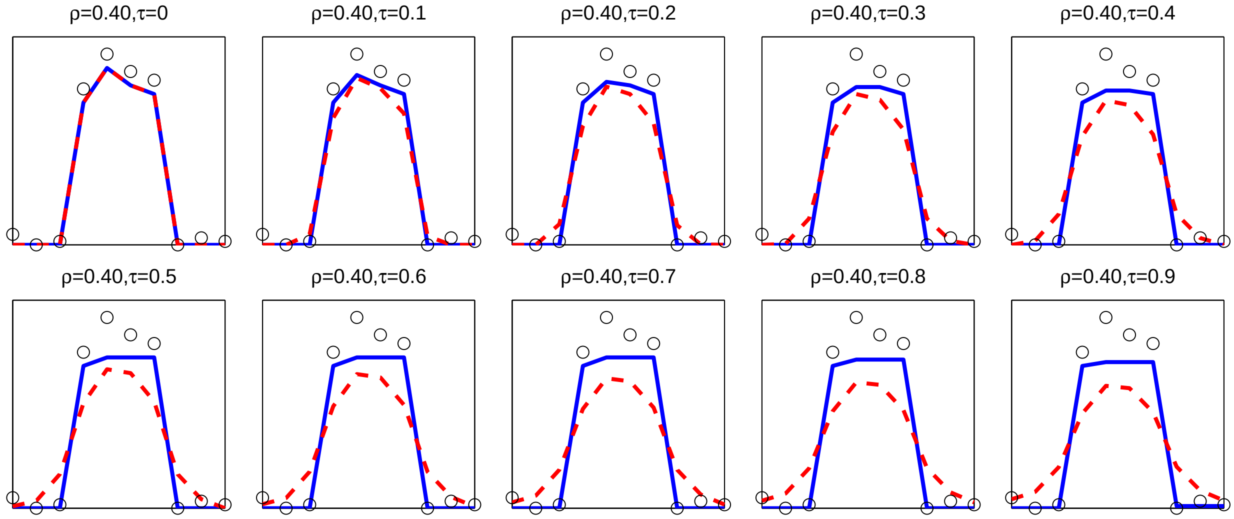


Figure A.1: Use of ℓ_1-norm penalty for local constancy (in blue solid line) versus squared ℓ_2-norm penalty for local constancy (in red dashed line) for resembling known values (black circles) with a sparse and locally constant profile. Note that the squared ℓ_2-norm penalty for differences does not produce a locally (piecewise) constant solution

Appendix B

Sherman-Woodbury-Morrison Formula

Algorithms 2.1, 3.1 and 4.1 update $\mathbf{W}^{-1}$ by using the Sherman-Woodbury-Morrison formula.

Note that when iterating from one variable to the next one, only one row and column change on matrix $\mathbf{W}$. Without loss of generality, let assume such row and column is the last one. The change in $\mathbf{W}$ due to the update of that row and column is denoted as $\mathbf{B}\mathbf{C}^{\mathrm{T}}$. The Sherman-Woodbury-Morrison formula for computing the inverse of the updated matrix $\mathbf{W}+\mathbf{B}\mathbf{C}^{\mathrm{T}}$ becomes:

$$(\mathbf{W}+\mathbf{B}\mathbf{C}^{\mathrm{T}})^{-1} = \mathbf{W}^{-1} - (\mathbf{W}^{-1}\mathbf{B})(\mathbf{I}+\mathbf{C}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{B})(\mathbf{C}^{\mathrm{T}}\mathbf{W}^{-1})$$

$$\mathbf{B} = \begin{bmatrix} \delta_1 & 0 \\ \delta_2 & 0 \\ \vdots & \vdots \\ \delta_{N-1} & 0 \\ \delta_N & 1 \end{bmatrix} \quad , \quad \mathbf{C} = \begin{bmatrix} 0 & \delta_1 \\ 0 & \delta_2 \\ \vdots & \vdots \\ 0 & \delta_{N-1} \\ 1 & \delta_N \end{bmatrix} \quad , \quad \mathbf{B}\mathbf{C}^{\mathrm{T}} = \begin{bmatrix} 0 & 0 & \cdots & 0 & \delta_1 \\ 0 & 0 & \cdots & 0 & \delta_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \delta_{N-1} \\ \delta_1 & \delta_2 & \cdots & \delta_{N-1} & 2\delta_N \end{bmatrix}$$

Appendix C

Technical Lemma for Variable Selection

In Theorem 3.2, we use four matrix norm inequalities that are less common in the literature. In this section, we prove them in detail.

Lemma C.1. *For* $\mathbf{A} \in \mathbb{R}^{N \times N}$*, the following conditions hold:*

$$\begin{array}{l} \text{i.}\ \ \|\mathbf{A}\|_2 \leq \sqrt{N}\|\mathbf{A}\|_{\infty,2} \\ \text{ii.}\ \ \|\mathbf{A}\|_2 \leq N\|\mathbf{A}\|_{\infty,1} \\ \text{iii.}\ \|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_{1,2} \\ \text{iv.}\ \mathbf{A} \succ \mathbf{0} \Rightarrow \|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_{1,\infty} \end{array} \tag{C.1}$$

Proof. Claim i follows from $\|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_{\mathfrak{F}} \leq \sqrt{N}\|\mathbf{A}\|_{\infty,2}$. The last inequality is equivalent to $\|\mathbf{A}\|_{\mathfrak{F}}^2 \leq N\|\mathbf{A}\|_{\infty,2}^2 \Rightarrow \sum_{n_1 n_2} a_{n_1 n_2}^2 \leq N \max_{n_1} \left(\sum_{n_2} a_{n_1 n_2}^2\right)$. Let $|c_{n_1}| = \sum_{n_2} a_{n_1 n_2}^2$, we get $\sum_{n_1} |c_{n_1}| \leq N \max_{n_1} |c_{n_1}|$. This is equivalent to $\|\mathbf{c}\|_1 \leq N\|\mathbf{c}\|_\infty$, and we prove our claim.

Claim ii follows from $\|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_{\mathfrak{F}} \leq \|\mathbf{A}\|_1 \leq N\|\mathbf{A}\|_{\infty,1}$. The last inequality is equivalent to $\sum_{n_1 n_2} |a_{n_1 n_2}| \leq N \max_{n_1} \left(\sum_{n_2} |a_{n_1 n_2}|\right)$. Let $|c_{n_1}| = \sum_{n_2} |a_{n_1 n_2}|$, we get $\sum_{n_1} |c_{n_1}| \leq N \max_{n_1} |c_{n_1}|$. This is equivalent to $\|\mathbf{c}\|_1 \leq N\|\mathbf{c}\|_\infty$, and we prove our claim.

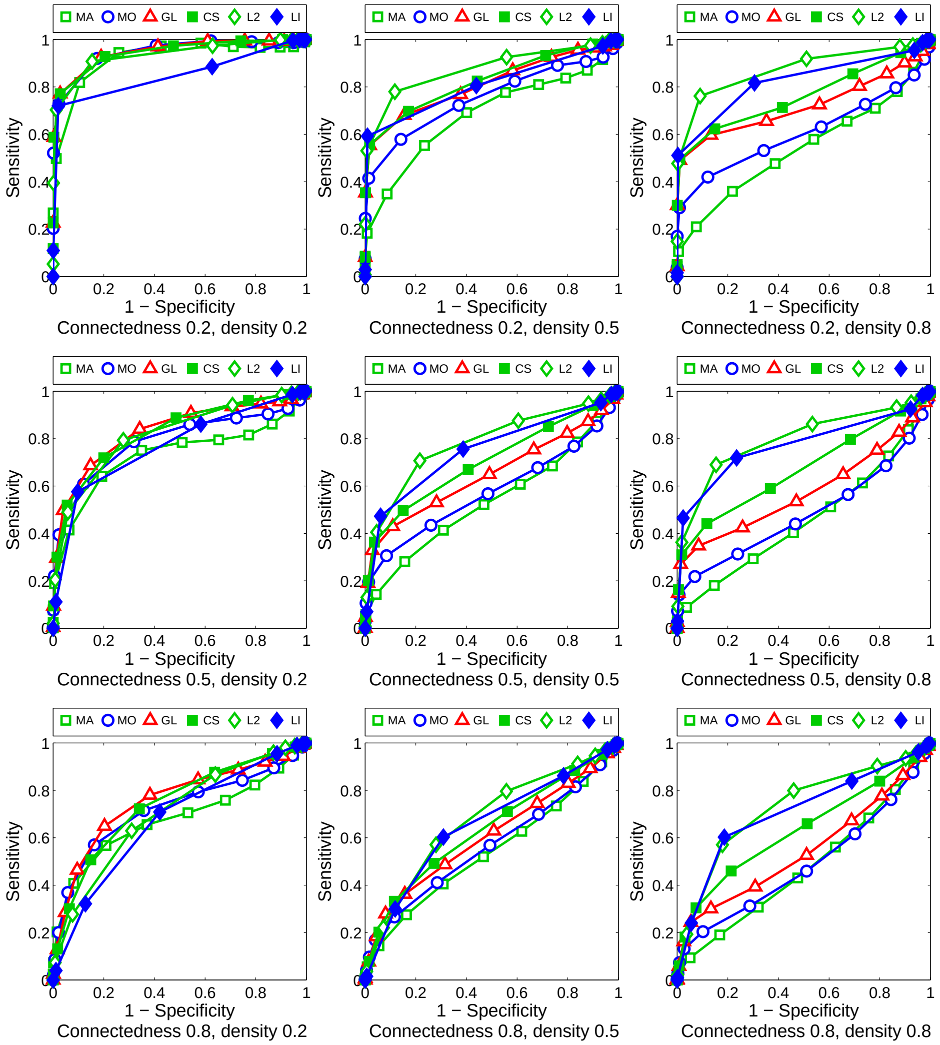
Claim iii follows from $\|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_{\mathfrak{F}} \leq \|\mathbf{A}\|_{1,2}$. The last inequality is equivalent to $\sqrt{\sum_{n_1 n_2} a_{n_1 n_2}^2} \leq \sum_{n_1} \sqrt{\sum_{n_2} a_{n_1 n_2}^2}$. Let $c_{n_1}^2 = \sum_{n_2} a_{n_1 n_2}^2$, we get $\sqrt{\sum_{n_1} c_{n_1}^2} \leq \sum_{n_1} \sqrt{c_{n_1}^2} = \sum_{n_1} |c_{n_1}|$. This is equivalent to $\|\mathbf{c}\|_2 \leq \|\mathbf{c}\|_1$, and we prove our claim.

Claim iv further assumes that $\mathbf{A}$ is symmetric and positive definite. In this case the spectral radius is less than or equal to any induced norm, specifically the $\ell_{\infty,1}$-norm also called the *max absolute row sum* norm. The inequality we want to prove is $\|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_{\infty,1} \leq \|\mathbf{A}\|_{1,\infty}$. The last inequality is equivalent to $\max_{n_1} \left(\sum_{n_2} |a_{n_1 n_2}|\right) \leq \sum_{n_1} \left(\max_{n_2} |a_{n_1 n_2}|\right)$, which follows from the Jensen's inequality. □

Appendix D

Additional Experimental Results for Variable Selection

In what follows, we test the performance of our methods with respect to edge density and the proportion of connected nodes. The following results complement Figure 3.2 which reported KL divergence between the recovered models and the ground truth for the “low variance confounders” regime. Figures D.1 and D.2 show the ROC curves and KL divergence between the recovered models and the ground truth for the “high variance confounders” regime. Our $\ell_{1,2}$ and $\ell_{1,\infty}$ methods recover ground truth edges better than competing methods (higher ROC) when edge density among connected nodes is moderate (0.5) to high (0.8), regardless of the proportion of connected nodes. Our proposed methods get similarly good probability distributions (comparable KL divergence) than the other techniques. In the “low variance confounders” regime reported in Figures D.3 and 3.2, our proposed methods produce better probability distributions (lower KL divergence) than the remaining techniques. The behavior of the ROC curves is similar to the “high variance confounders” regime.



Connectedness 0.2, density 0.2

Connectedness 0.2, density 0.5

Connectedness 0.2, density 0.8

Connectedness 0.5, density 0.2

Connectedness 0.5, density 0.5

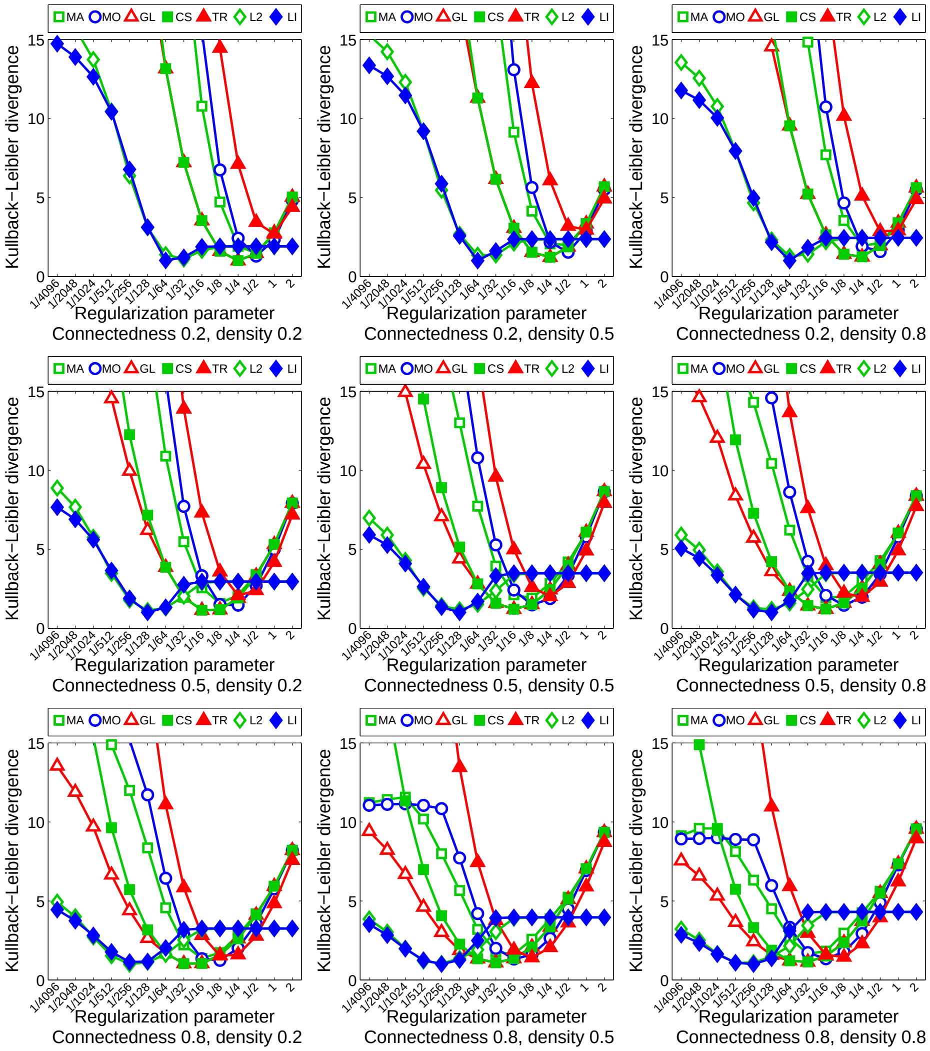
Connectedness 0.5, density 0.8

Connectedness 0.8, density 0.2

Connectedness 0.8, density 0.5

Connectedness 0.8, density 0.8

Figure D.1: ROC curves for structures learnt for the "high variance confounders" regime ($N = 50$ variables, different connectedness and density levels). Our proposed methods $\ell_{1,2}$ (L2) and $\ell_{1,\infty}$ (LI) recover the ground truth edges better than Meinshausen-Bühlmann with AND-rule (MA), OR-rule (MO), graphical lasso (GL) and covariance selection (CS), when the edge density among the connected nodes is moderate (center) to high (right).



Connectedness 0.2, density 0.2

Connectedness 0.2, density 0.5

Connectedness 0.2, density 0.8

Connectedness 0.5, density 0.2

Connectedness 0.5, density 0.5

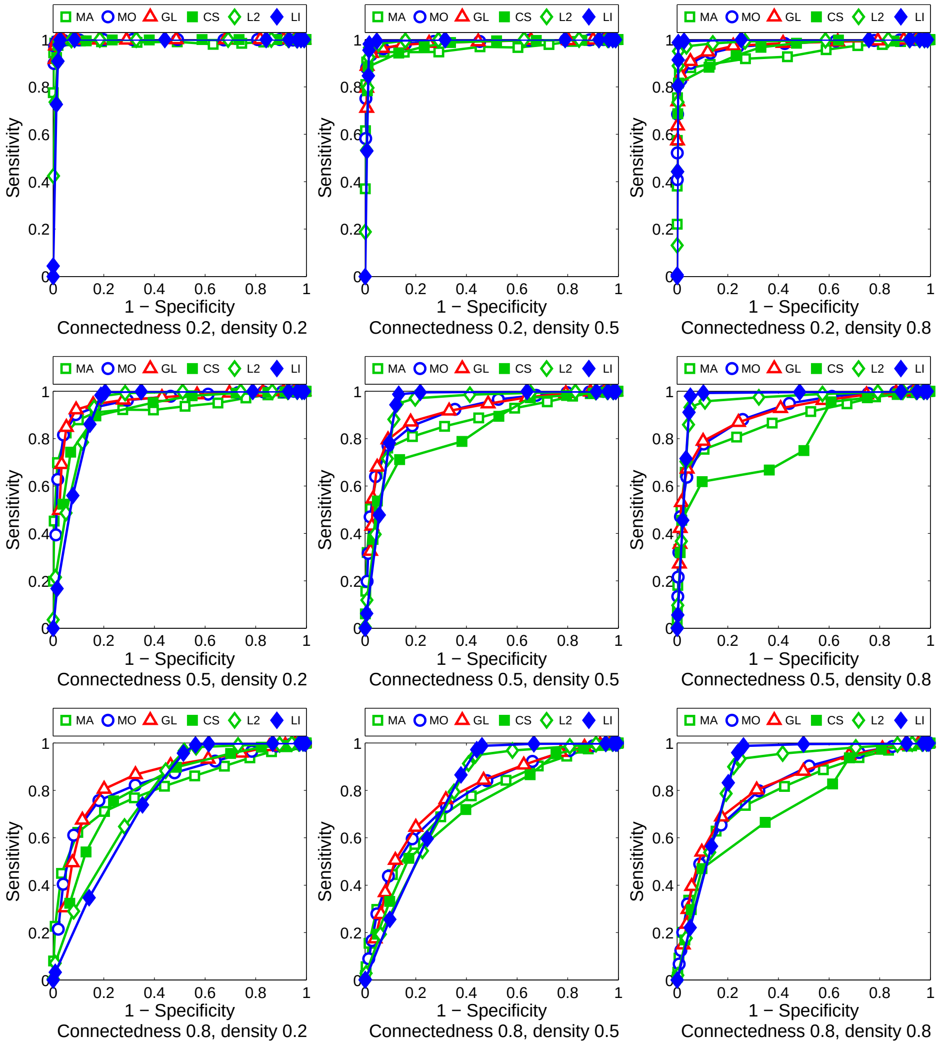
Connectedness 0.5, density 0.8

Connectedness 0.8, density 0.2

Connectedness 0.8, density 0.5

Connectedness 0.8, density 0.8

Figure D.2: Cross-validated KL divergence for structures learnt for the "high variance confounders" regime ($N = 50$ variables, different connectedness and density levels). Our proposed methods $\ell_{1,2}$ (L2) and $\ell_{1,\infty}$ (LI) produce similarly good probability distributions than Meinshausen-Bühlmann with AND-rule (MA), OR-rule (MO), graphical lasso (GL), covariance selection (CS) and Tikhonov regularization (TR).



Connectedness 0.2, density 0.2

Connectedness 0.2, density 0.5

Connectedness 0.2, density 0.8

Connectedness 0.5, density 0.2

Connectedness 0.5, density 0.5

Connectedness 0.5, density 0.8

Connectedness 0.8, density 0.2

Connectedness 0.8, density 0.5

Connectedness 0.8, density 0.8

Figure D.3: ROC curves for structures learnt for the "low variance confounders" regime ($N = 50$ variables, different connectedness and density levels). Our proposed methods $\ell_{1,2}$ (L2) and $\ell_{1,\infty}$ (LI) recover the ground truth edges better than Meinshausen-Bühlmann with AND-rule (MA), OR-rule (MO), graphical lasso (GL) and covariance selection (CS), when the edge density among the connected nodes is moderate (center) to high (right).

Appendix E

List of Brain Regions for Multi-Task Learning Experiments

Next, we present the 157 regions used in Section 4.9. In order to not make the list unnecessarily long, we use regular expressions, e.g. "(Left | Right) Amygdala" indicates that we used two regions: "Left Amygdala" and "Right Amygdala".

- Cerebellum: Cerebellar Lingual
- Cerebellum: (Culmen | Declive | Pyramis | Tuber | Uvula) of Vermis
- Cerebellum: (Left | Right) (Cerebellar Tonsil | Culmen | Declive | Dentate | Fastigium | Inferior Semi-Lunar Lobule | Nodule | Pyramis | Tuber | Uvula)
- Cerebrum: Hypothalamus
- Cerebrum: (Left | Right) (Amygdala | Claustrum | Hippocampus | Pulvinar | Putamen)
- Cerebrum: (Left | Right) (Anterior | Lateral Dorsal | Lateral Posterior | Medial Dorsal | Midline | Ventral Anterior | Ventral Lateral | Ventral Posterior Lateral | Ventral Posterior Medial) Nucleus
- Cerebrum: (Left | Right) Brodmann area (1 | 2 | . . . | 47)
- Cerebrum: (Left | Right) Caudate (Body | Head | Tail)
- Cerebrum: (Left | Right) (Lateral | Medial) Globus Pallidus
- Brainstem: (Left | Right) (Mammillary Body | Red Nucleus | Substantia Nigra | Subthalamic Nucleus)

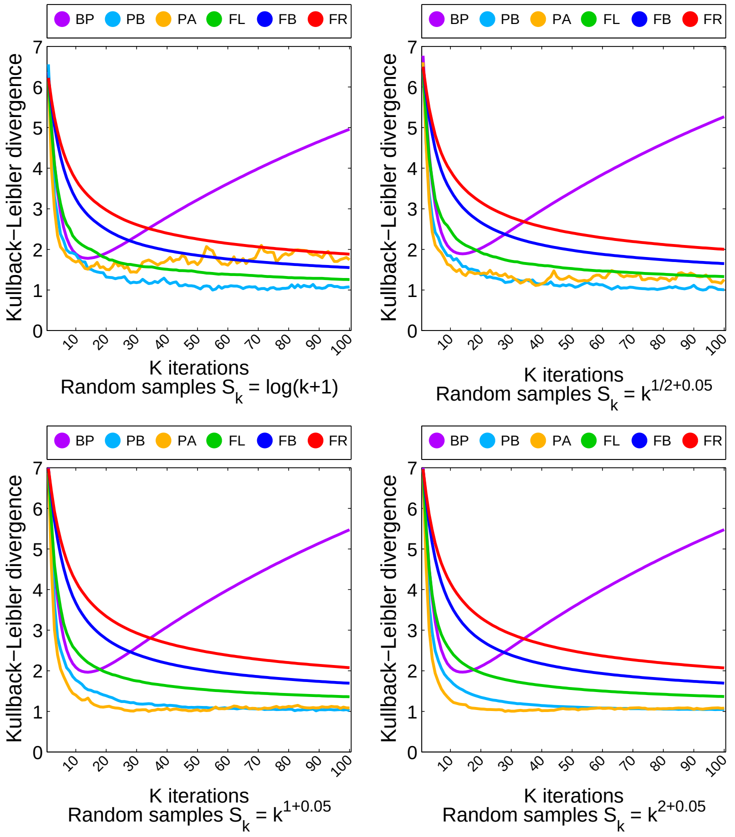
Appendix F

Additional Experimental Results for Discrete MRFs

First, we complement the results in Figure 5.1. We show the Kullback-Leibler divergence to the ground truth in Figure F.1.

Note that we assumed a "zero field" regime for Figures 5.1 and F.1 where $\mathbf{b}_g = \mathbf{0}$. We also report results in Figures F.2 and F.3 for the "non-zero field" regime where each entry of $\mathbf{b}_g$ is generated uniformly at random from $[-1; +1]$.

We also evaluate a "mean field sampler" by first finding the mean field distribution and then performing importance sampling with the mean field trial. We report results for the "zero field" regime in Figures F.4 and F.5, and for the "non-zero field" regime in Figures F.6 and F.7.



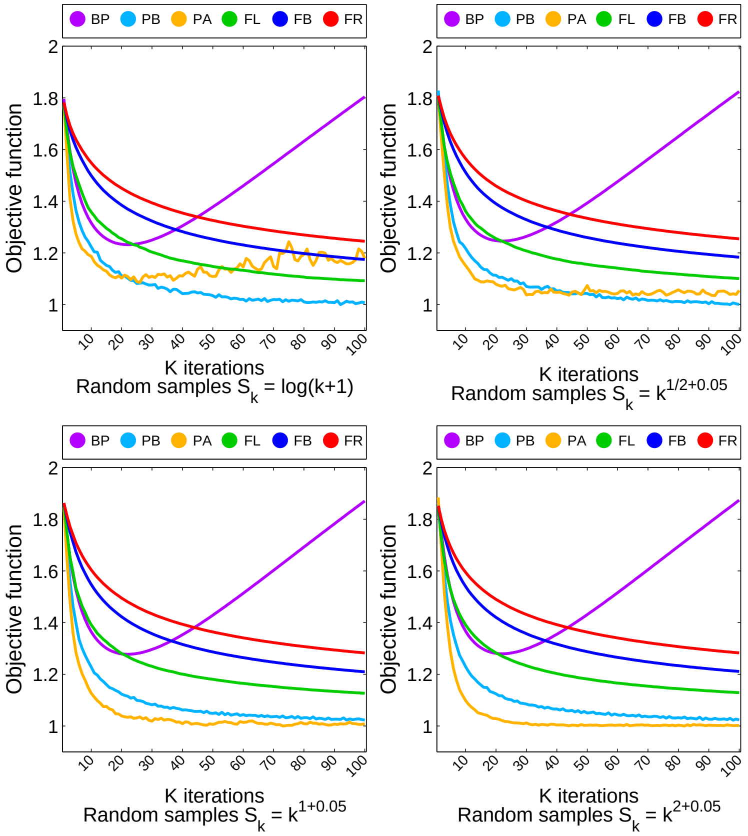
Random samples $S_k = \log(k+1)$

Random samples $S_k = k^{1/2+0.05}$

Random samples $S_k = k^{1+0.05}$

Random samples $S_k = k^{2+0.05}$

Figure F.1: Kullback-Leibler divergence to the ground truth for different settings of increasing number of random samples for the “zero-field” regime and “Gibbs sampler”. Basic (PB) and accelerated (PA) are noisier and require more samples than last point (FL), basic (FB) and robust (FR) forward-backward splitting in order to generalize well, but they exhibit faster convergence. Belief propagation (BP) does not generalize well.



Random samples $S_k = \log(k+1)$

Random samples $S_k = k^{1/2+0.05}$

Random samples $S_k = k^{1+0.05}$

Random samples $S_k = k^{2+0.05}$

Figure F.2: Objective function for different settings of increasing number of random samples for the "non-zero field" regime and "Gibbs sampler". Basic (PB) and accelerated (PA) are noisier and require more samples than last point (FL), basic (FB) and robust (FR) forward-backward splitting in order to converge, but they exhibit faster convergence. Belief propagation (BP) does not converge.

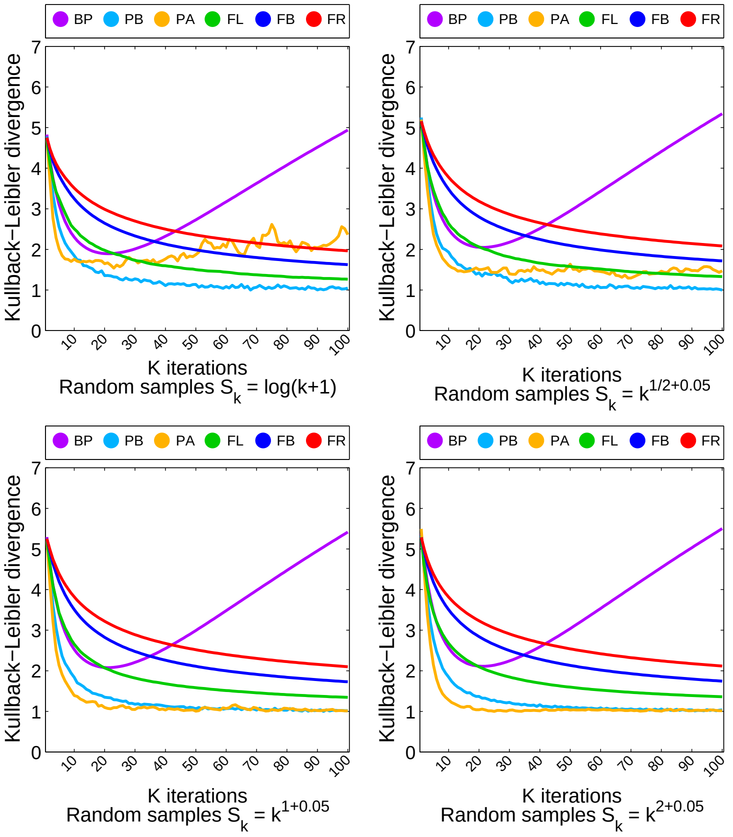
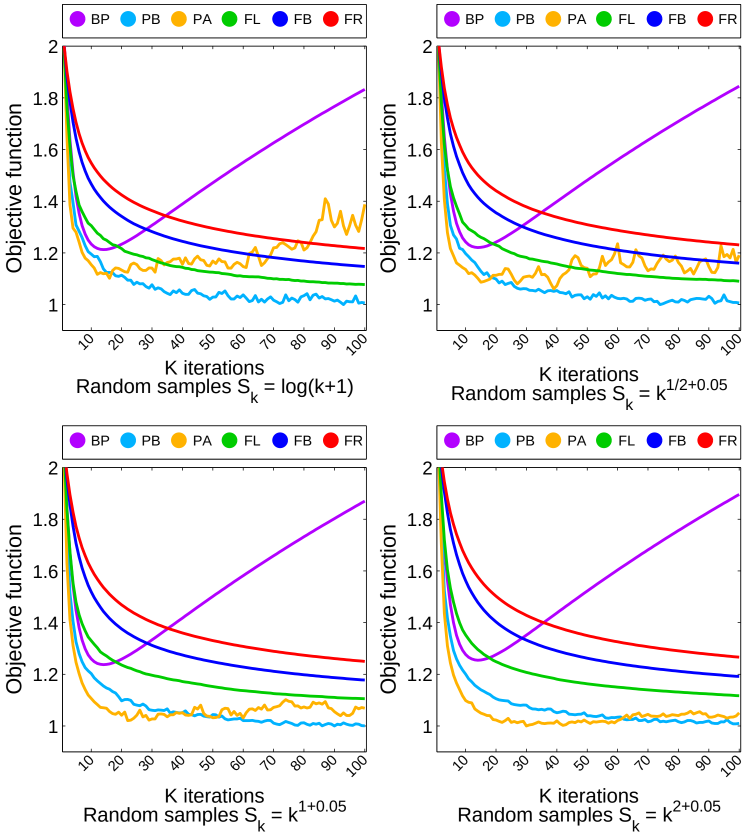


Figure F.3: Kullback-Leibler divergence to the ground truth for different settings of increasing number of random samples for the "non-zero field" regime and "Gibbs sampler". Basic (PB) and accelerated (PA) are noisier and require more samples than last point (FL), basic (FB) and robust (FR) forward-backward splitting in order to generalize well, but they exhibit faster convergence. Belief propagation (BP) does not generalize well.



Random samples $S_k = \log(k+1)$

Random samples $S_k = k^{1/2+0.05}$

Random samples $S_k = k^{1+0.05}$

Random samples $S_k = k^{2+0.05}$

Figure F.4: Objective function for different settings of increasing number of random samples for the “zero-field” regime and “mean field sampler”. Basic (PB) and accelerated (PA) are noisier and require more samples than last point (FL), basic (FB) and robust (FR) forward-backward splitting in order to converge, but they exhibit faster convergence. Belief propagation (BP) does not converge.

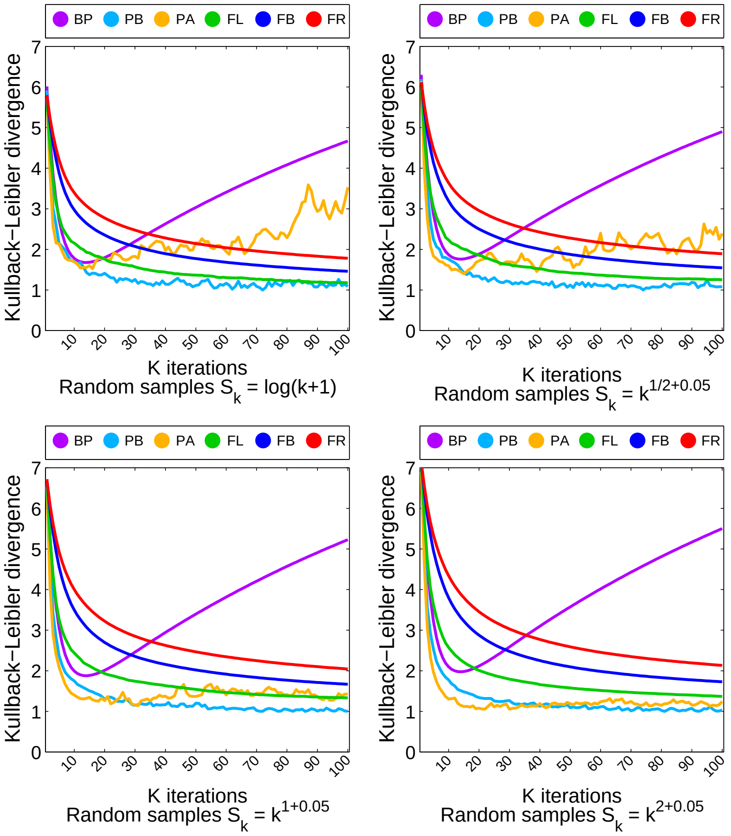
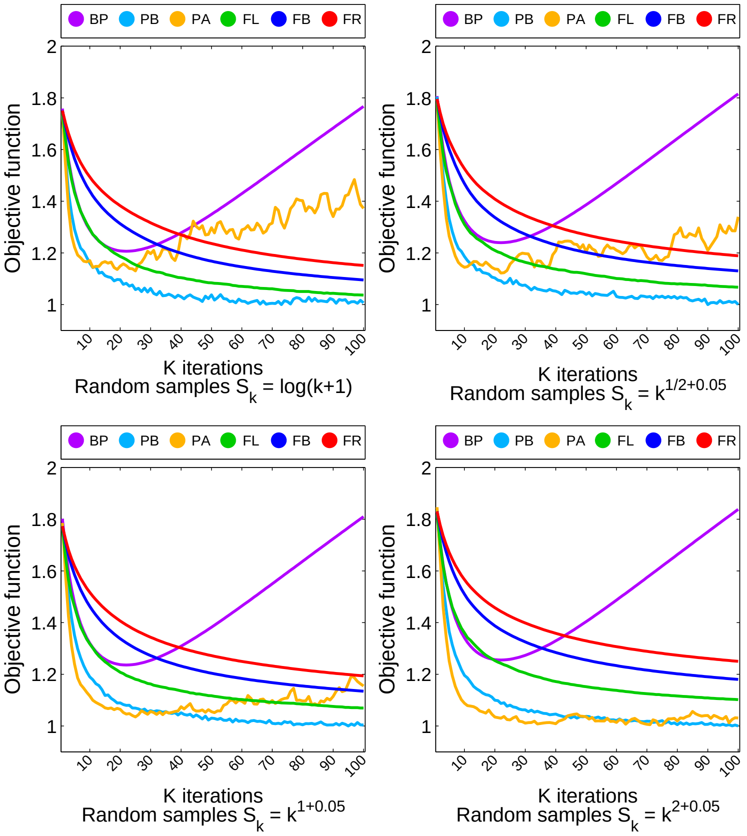


Figure F.5: Kullback-Leibler divergence to the ground truth for different settings of increasing number of random samples for the “zero-field” regime and “mean field sampler”. Basic (PB) and accelerated (PA) are noisier and require more samples than last point (FL), basic (FB) and robust (FR) forward-backward splitting in order to generalize well, but they exhibit faster convergence. Belief propagation (BP) does not generalize well.



Random samples $S_k = \log(k+1)$

Random samples $S_k = k^{1/2+0.05}$

Random samples $S_k = k^{1+0.05}$

Random samples $S_k = k^{2+0.05}$

Figure F.6: Objective function for different settings of increasing number of random samples for the "non-zero field" regime and "mean field sampler". Basic (PB) and accelerated (PA) are noisier and require more samples than last point (FL), basic (FB) and robust (FR) forward-backward splitting in order to converge, but they exhibit faster convergence. Belief propagation (BP) does not converge.

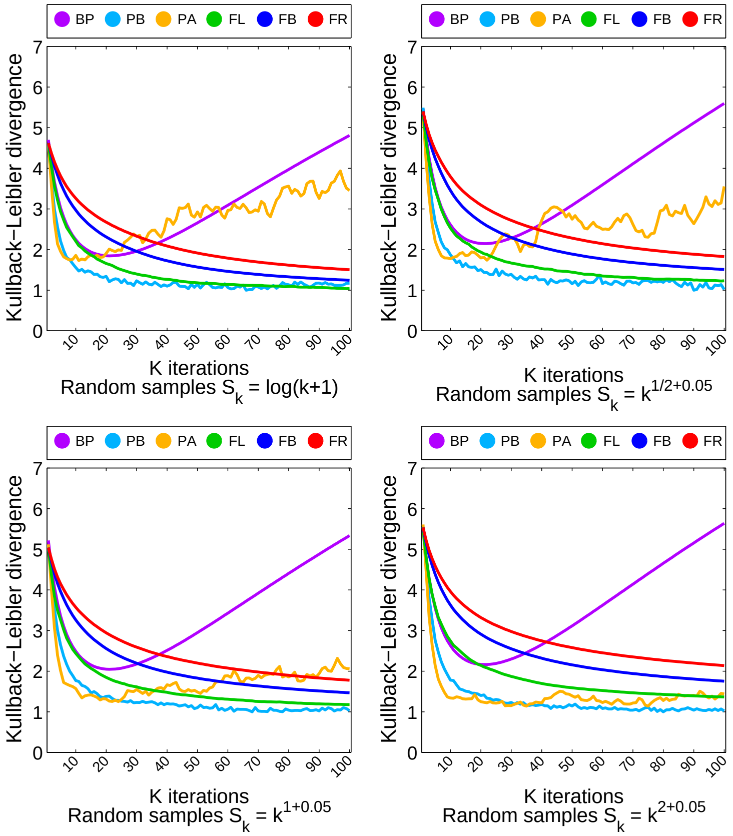


Figure F.7: Kullback-Leibler divergence to the ground truth for different settings of increasing number of random samples for the "non-zero field" regime and "mean field sampler". Basic (PB) and accelerated (PA) are noisier and require more samples than last point (FL), basic (FB) and robust (FR) forward-backward splitting in order to generalize well, but they exhibit faster convergence. Belief propagation (BP) does not generalize well.

Appendix G

Norm Bound for Linear Regression

We show that the weight vector $\mathbf{w}$ of the linear regression has bounded norm, i.e. $\|\mathbf{w}\|_2 \leq \beta$.

Given the dependent variable for T samples $\mathbf{y} \in \mathbb{R}^T$ and the matrix of N regressors for each of the T samples $\mathbf{X} \in \mathbb{R}^{N \times T}$, the Tikhonov regularized linear regression problem is given by:

$$\min_{\mathbf{w} \in \mathbb{R}^N} \|\mathbf{y} - \mathbf{w}^{\mathrm{T}}\mathbf{X}\|_2^2 + \rho\|\mathbf{w}\|_2^2 \tag{G.1}$$

for $\rho > 0$. It is well known [Boyd and Vandenberghe, 2006] that the optimal solution of the above problem is $\mathbf{w}^* = (\mathbf{X}\mathbf{X}^{\mathrm{T}} + \rho\mathbf{I})^{-1}\mathbf{X}\mathbf{y}$. Let $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}}$ be the singular value decomposition of $\mathbf{X}$, where $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$, $\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$ and $\mathbf{D}$ is a diagonal matrix with diagonal entries $(\forall n)$ $d_n \geq 0$ (i.e. the non-negative singular values). Then the optimal solution can be written as $\mathbf{w}^* = \mathbf{V}\mathbf{D}'\mathbf{U}^{\mathrm{T}}\mathbf{y}$, where $\mathbf{D}'$ is a diagonal matrix with diagonal entries $(\forall n)$ $\frac{d_n}{d_n^2+\rho} > 0$.

In order to find an upper bound for $\|\mathbf{D}'\|_2$, we need to find the maximum possible value of $f(d_n) = \frac{d_n}{d_n^2+\rho}$. By deriving with respect to d_n we can find that the optimal value is $d_n^* = \sqrt{\rho}$ and therefore $f(d_n^*) = \frac{1}{1+\sqrt{\rho}}$. Finally, $\|\mathbf{w}^*\|_2 \leq \|\mathbf{V}\|_2\|\mathbf{D}'\|_2\|\mathbf{U}\|_2\|\mathbf{y}\|_2 \leq \|\mathbf{D}'\|_2\|\mathbf{y}\|_2 \leq \frac{1}{1+\sqrt{\rho}}\|\mathbf{y}\|_2 = \beta$.

Appendix H

Norm Bound for Gaussian MRFs

We show that the precision matrix $\mathbf{\Omega}$ has bounded norm, i.e. $\alpha\mathbf{I} \preceq \mathbf{\Omega} \preceq \beta\mathbf{I}$. Similarly, since the covariance matrix is the inverse of the precision matrix, we show that the covariance matrix $\mathbf{\Sigma}$ has bounded norm, i.e. $\frac{1}{\beta}\mathbf{I} \preceq \mathbf{\Sigma} \preceq \frac{1}{\alpha}\mathbf{I}$.

Given a dense sample covariance matrix $\widehat{\mathbf{\Sigma}} \succeq \mathbf{0}$, consider the Tikhonov regularized precision matrix $\mathbf{\Omega} = (\widehat{\mathbf{\Sigma}} + \rho\mathbf{I})^{-1}$. Note that the minimum eigenvalue of $\widehat{\mathbf{\Sigma}} + \rho\mathbf{I}$ is ρ and the maximum eigenvalue is $\|\widehat{\mathbf{\Sigma}}\| + \rho$. Therefore, $\alpha = \frac{1}{\|\widehat{\mathbf{\Sigma}}\|+\rho}$ and $\beta = \frac{1}{\rho}$.

Similar bounds can be obtained for sparseness promoting (ℓ_1) methods. The problem of finding a sparse precision matrix $\mathbf{\Omega}$ by regularized maximum likelihood estimation is given by:

$$\max_{\mathbf{\Omega}\succ\mathbf{0}} \Big(\log\det\mathbf{\Omega} - \langle\widehat{\mathbf{\Sigma}}, \mathbf{\Omega}\rangle - \rho\|\mathbf{\Omega}\|_1\Big) \tag{H.1}$$

for $\rho > 0$. Banerjee et al. [2006] proved that the optimal solution to the above problem is bounded by $\alpha = \frac{1}{\|\widehat{\mathbf{\Sigma}}\|_2+N\rho}$ and $\beta = \frac{N}{\rho}$.

Appendix I

Loose Kullback-Leibler Bound for Gaussian MRFs

If we use Lemma 6.21 for factor graphs, we will obtain a loose bound of the Kullback-Leibler divergence for Gaussian graphical models. More specifically $\mathbb{E}_{\mathcal{P}}[\|\boldsymbol{\psi}(\mathbf{x})\|_p]$ for $\boldsymbol{\psi}(\mathbf{x}) = \mathbf{vec}(\mathbf{x}\mathbf{x}^{\mathrm{T}})$ and $p = 2$ becomes $\mathbb{E}_{\mathcal{P}}[\|\mathbf{x}\|_2^2] = \mathbb{E}_{\mathcal{P}}[\mathbf{x}^{\mathrm{T}}\mathbf{x}] = \int_{\mathbf{x}} \frac{(\det \boldsymbol{\Omega})^{1/2}}{(2\pi)^{N/2}} e^{-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Omega}\mathbf{x}} \mathbf{x}^{\mathrm{T}}\mathbf{x} \equiv B$. Since $\det \boldsymbol{\Omega} \leq \beta^N$ and $(\forall \mathbf{x})\ \mathbf{x}^{\mathrm{T}}\boldsymbol{\Omega}\mathbf{x} \geq \alpha \mathbf{x}^{\mathrm{T}}\mathbf{x}$, we have $B \leq \frac{\beta^{N/2}}{(2\pi)^{N/2}} \int_{\mathbf{x}} e^{-\frac{\alpha}{2}\mathbf{x}^{\mathrm{T}}\mathbf{x}} \mathbf{x}^{\mathrm{T}}\mathbf{x} = \frac{N\beta^{N/2}}{\alpha^{N/2+1}}$.

Finally, the bound in Lemma 6.21 becomes $\overline{K} = \frac{2N\beta^{N/2}}{\alpha^{N/2+1}}$.

Appendix J

Parametrization of Gaussian MRFs by Covariance Matrices

In Section 6.4, we analyzed parametrization of Gaussian graphical models by using precision matrices. Here, we also analyze parametrization by using covariance matrices. Similarly, since the covariance matrix is the inverse of the precision matrix, we assume that the covariance matrix $\mathbf{\Sigma}$ has bounded norm, i.e. $\frac{1}{\beta}\mathbf{I} \preceq \mathbf{\Sigma} \preceq \frac{1}{\alpha}\mathbf{I}$ or equivalently $\|\mathbf{\Sigma}^{-1}\|_2 \leq \beta$ and $\|\mathbf{\Sigma}\|_2 \leq \frac{1}{\alpha}$.

Lemma J.1. *Given the covariance matrix $\mathbf{\Sigma} \succ \mathbf{0}$, the Gaussian graphical model parameterized by $\mathbf{\Theta} = \mathbf{\Sigma}$, $\mathbf{\Sigma} \in \mathbb{R}^{N\times N}$ with probability density function:*

$$p(\mathbf{x}|\mathbf{\Theta}) = \frac{1}{(2\pi)^{N/2}(\det \mathbf{\Sigma})^{1/2}} e^{-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{x}} \tag{J.1}$$

is $(\ell_2, \frac{\beta^2\|\mathbf{x}\|_2^2}{2} + \frac{\beta}{2})$-Lipschitz continuous.

Proof. Let $f(\mathbf{\Sigma}) = \log p(\mathbf{x}|\mathbf{\Theta}) = \frac{1}{2}(-\log\det\mathbf{\Sigma} - N\log(2\pi) - \mathbf{x}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{x})$. By deriving $\partial f/\partial \mathbf{\Sigma} = \frac{1}{2}(-\mathbf{\Sigma}^{-1} + \mathbf{\Sigma}^{-1}\mathbf{x}\mathbf{x}^{\mathrm{T}}\mathbf{\Sigma}^{-1})$. Therefore $\|\partial f/\partial \mathbf{\Omega}\|_2 \leq \frac{1}{2}(\|\mathbf{\Sigma}^{-1}\|_2 + \|\mathbf{\Sigma}^{-1}\|_2\|\mathbf{x}\mathbf{x}^{\mathrm{T}}\|_2\|\mathbf{\Sigma}^{-1}\|_2) = \frac{1}{2}(\|\mathbf{\Sigma}^{-1}\|_2 + \|\mathbf{\Sigma}^{-1}\|_2^2\|\mathbf{x}\|_2^2) \leq \frac{1}{2}(\beta + \beta^2\|\mathbf{x}\|_2^2)$. By Definitions 6.4 and 6.5, we prove our claim. □

Lemma J.2. *Given two Gaussian graphical models parameterized by their covariance matrices as in eq.*(J.1)*, i.e. $\mathcal{P}_1 = p(\cdot|\mathbf{\Sigma}_1)$ and $\mathcal{P}_2 = p(\cdot|\mathbf{\Sigma}_2)$, the Kullback-Leibler divergence from $\mathcal{P}_1$ to $\mathcal{P}_2$:*

$$\mathcal{KL}(\mathcal{P}_1||\mathcal{P}_2) = \frac{1}{2}\left(\log\frac{\det\mathbf{\Sigma}_2}{\det\mathbf{\Sigma}_1} + \langle\mathbf{\Sigma}_1, \mathbf{\Sigma}_2^{-1}\rangle - N\right) \tag{J.2}$$

is bounded as follows:

$$\mathcal{KL}(\mathcal{P}_1||\mathcal{P}_2) \leq \beta\|\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2\|_2 \tag{J.3}$$

Proof. First, we show that $f(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2) = \mathcal{KL}(\mathcal{P}_1||\mathcal{P}_2)$ is Lipschitz continuous with respect to $\mathbf{\Sigma}_1$. By deriving $\partial f/\partial\mathbf{\Sigma}_1 = \frac{1}{2}(-\mathbf{\Sigma}_1^{-1} + \mathbf{\Sigma}_2^{-1})$. Therefore $\|\partial f/\partial\mathbf{\Sigma}_1\|_2 \leq \frac{1}{2}(\|\mathbf{\Sigma}_1^{-1}\|_2 + \|\mathbf{\Sigma}_2^{-1}\|_2) \leq \frac{1}{2}(\beta + \beta) = \beta$.

Second, since f is Lipschitz continuous with respect to its first parameter, we have $(\forall\mathbf{\Sigma})\ |f(\mathbf{\Sigma}_1, \mathbf{\Sigma}) - f(\mathbf{\Sigma}_2, \mathbf{\Sigma})| \leq \beta\|\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2\|_2$. In particular, let $\mathbf{\Sigma} = \mathbf{\Sigma}_2$ and since $f(\mathbf{\Sigma}_2, \mathbf{\Sigma}_2) = 0$ and $|f(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2)| = f(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2)$ by properties of the Kullback-Leibler divergence, we prove our claim. □

Appendix K

Negative Results for Linear Influence Games

Counting the Number of Equilibria is NP-hard.

Here we provide a proof that establishes NP-hardness of counting the number of Nash equilibria, and thus also of evaluating the log-likelihood function for our generative model. A #P-hardness proof was originally provided by Irfan and Ortiz [2011], here we present a related proof for completeness. The reduction is from the *set partition problem* for a specific instance of a single *non-absolutely-indifferent* player.

Recall the *set partition problem*: given a multiset of n positive numbers $\{a_1, \ldots, a_n\}$, SetPartition($\mathbf{a}$) answers "yes" if and only if it is possible to partition the numbers into two disjoint subsets $\mathcal{S}_1$ and $\mathcal{S}_2$ such that $\mathcal{S}_1 \cap \mathcal{S}_2 = \emptyset$, $\mathcal{S}_1 \cup \mathcal{S}_2 = \{1, \ldots, n\}$ and $\sum_{i \in \mathcal{S}_1} a_i - \sum_{i \in \mathcal{S}_2} a_i = 0$; otherwise it answers "no". The set partition problem is equivalent to the *subset sum problem*, in which given a set of positive numbers $\{a_1, \ldots, a_n\}$ and a target sum $c > 0$, SubSetSum($\mathbf{a}, c$) answers "yes" if and only if there is a subset $\mathcal{S} \subset \{1, \ldots, n\}$ such that $\sum_{i \in \mathcal{S}} a_i = c$; otherwise it answers "no". The equivalence between set partition and subset sum follows from SetPartition($\mathbf{a}$) = SubSetSum($\mathbf{a}, \frac{1}{2} \sum_i a_i$).

For clarity of exposition, we drop the subindices in the following lemma. Let $\mathbf{w} \equiv \mathbf{w}_{i,-i} \in \mathbb{R}^{n-1}$ and $b \equiv b_i \in \mathbb{R}$.

Lemma K.1. *The problem of counting Nash equilibria considered in Claim ii of Lemma 7.20 reduces to the set partition problem. More specifically, given* $(\forall i)\ w_i > 0, b = 0$, *answering whether* $\sum_{\mathbf{x}} 1[\mathbf{w}^{\mathrm{T}}\mathbf{x} - b = 0] > 0$ *is equivalent to answering* SetPartition($\mathbf{w}$).

Proof. Let $\mathcal{S}_1(\mathbf{x}) = \{i | x_i = +1\}$ and $\mathcal{S}_2(\mathbf{x}) = \{i | x_i = -1\}$. We can rewrite $\sum_{\mathbf{x}} 1[\mathbf{w}^{\mathrm{T}}\mathbf{x} - b = 0]$ as a sum of *set partition* conditions, i.e. $\sum_{\mathbf{x}} 1[\sum_{i \in \mathcal{S}_1(\mathbf{x})} w_i - \sum_{i \in \mathcal{S}_2(\mathbf{x})} w_i = 0]$. Therefore, if no tuple $\mathbf{x}$ fulfills the condition, the sum is zero and SetPartition($\mathbf{w}$) answers "no". On the other hand, if at least one tuple $\mathbf{x}$ fulfills the condition, the sum is greater than zero and SetPartition($\mathbf{w}$) answers "yes". □

Computing the Pseudo-Likelihood is NP-hard.

We show that evaluating the pseudo-likelihood function for our generative model is NP-hard. First, consider a non-trivial influence game $\mathcal{G}$ in which eq.(7.3) simplifies to $p_{(\mathcal{G},q)}(\mathbf{x}) = q\frac{1[\mathbf{x}\in\mathcal{NE}(\mathcal{G})]}{|\mathcal{NE}(\mathcal{G})|} + (1-q)\frac{1[\mathbf{x}\notin\mathcal{NE}(\mathcal{G})]}{2^n-|\mathcal{NE}(\mathcal{G})|}$. Furthermore, assume the game $\mathcal{G} = (\mathbf{W}, \mathbf{b})$ has a single *non-absolutely-indifferent* player i and *absolutely-indifferent* players $\forall j \neq i$. Let $f_i(\mathbf{x}_{-i}) \equiv \mathbf{w}_{i,-i}{}^{\mathrm{T}}\mathbf{x}_{-i} - b_i$. By Claim i of Lemma 7.20, we have $1[\mathbf{x} \in \mathcal{NE}(\mathcal{G})] = 1[x_i f_i(\mathbf{x}_{-i}) \geq 0]$ and therefore $p_{(\mathcal{G},q)}(\mathbf{x}) = q\frac{1[x_i f_i(\mathbf{x}_{-i})\geq 0]}{|\mathcal{NE}(\mathcal{G})|} + (1-q)\frac{1-1[x_i f_i(\mathbf{x}_{-i})\geq 0]}{2^n-|\mathcal{NE}(\mathcal{G})|}$. Finally, by Lemma K.1 computing $|\mathcal{NE}(\mathcal{G})|$ is NP-hard even for this specific instance of a single *non-absolutely-indifferent* player.

Counting the Number of Equilibria is not (Lipschitz) Continuous.

We show that small changes in the parameters $\mathcal{G} = (\mathbf{W}, \mathbf{b})$ can produce big changes in $|\mathcal{NE}(\mathcal{G})|$. For instance, consider two games $\mathcal{G}_k = (\mathbf{W}_k, \mathbf{b}_k)$, where $\mathbf{W}_1 = \mathbf{0}, \mathbf{b}_1 = \mathbf{0}, |\mathcal{NE}(\mathcal{G}_1)| = 2^n$ and $\mathbf{W}_2 = \varepsilon(\mathbf{1}\mathbf{1}^{\mathrm{T}} - \mathbf{I}), \mathbf{b}_2 = \mathbf{0}, |\mathcal{NE}(\mathcal{G}_2)| = 2$ for $\varepsilon > 0$. For $\varepsilon \to 0$, any ℓ_p-norm $\|\mathbf{W}_1 - \mathbf{W}_2\|_p \to 0$ but $|\mathcal{NE}(\mathcal{G}_1)| - |\mathcal{NE}(\mathcal{G}_2)| = 2^n - 2$ remains constant.

The Log-Partition Function of an Ising Model is a Trivial Bound for Counting the Number of Equilibria.

Let $f_i(\mathbf{x}_{-i}) \equiv \mathbf{w}_{i,-i}{}^{\mathrm{T}}\mathbf{x}_{-i} - b_i$, $|\mathcal{NE}(\mathcal{G})| = \sum_{\mathbf{x}} \prod_i 1[x_i f_i(\mathbf{x}_{-i}) \geq 0] \leq \sum_{\mathbf{x}} \prod_i e^{x_i f_i(\mathbf{x}_{-i})} = \sum_{\mathbf{x}} e^{\mathbf{x}^{\mathrm{T}}\mathbf{W}\mathbf{x} - \mathbf{b}^{\mathrm{T}}\mathbf{x}} = \mathcal{Z}(\frac{1}{2}(\mathbf{W} + \mathbf{W}^{\mathrm{T}}), \mathbf{b})$, where $\mathcal{Z}$ denotes the partition function of an Ising model. Given convexity of $\mathcal{Z}$ [Koller and Friedman, 2009] and that the gradient vanishes at $\mathbf{W} = \mathbf{0}, \mathbf{b} = \mathbf{0}$, we know that $\mathcal{Z}(\frac{1}{2}(\mathbf{W} + \mathbf{W}^{\mathrm{T}}), \mathbf{b}) \geq 2^n$, which is the maximum $|\mathcal{NE}(\mathcal{G})|$.

Appendix L

Simultaneous Logistic Loss

Given that any loss $\ell(z)$ is a decreasing function, the following identity holds $\max_i \ell(z_i) = \ell(\min_i z_i)$. Hence, we can either upper-bound the max function by the logsumexp function or lower-bound the min function by a negative logsumexp. We chose the latter option for the logistic loss for the following reasons: Claim i of the following technical lemma shows that lower-bounding min generates a loss that is strictly less than upper-bounding max. Claim ii shows that lower-bounding min generates a loss that is strictly less than independently penalizing each player. Claim iii shows that there are some cases in which upper-bounding max generates a loss that is strictly greater than independently penalizing each player.

Lemma L.1. *For the logistic loss $\ell(z) = \log(1+e^{-z})$ and a set of $n > 1$ numbers $\{z_1, \ldots, z_n\}$:*

$$\begin{array}{ll} \text{i.} & (\forall z_1, \ldots, z_n)\ \max_i \ell(z_i) \le \ell\left(-\log \sum_i e^{-z_i}\right) < \log \sum_i e^{\ell(z_i)} \le \max_i \ell(z_i) + \log n \\ \text{ii.} & (\forall z_1, \ldots, z_n)\ \ell\left(-\log \sum_i e^{-z_i}\right) < \sum_i \ell(z_i) \\ \text{iii.} & (\exists z_1, \ldots, z_n)\ \log \sum_i e^{\ell(z_i)} > \sum_i \ell(z_i) \end{array} \tag{L.1}$$

Proof. Given a set of numbers $\{a_1, \ldots, a_n\}$, the max function is bounded by the logsumexp function by $\max_i a_i \le \log \sum_i e^{a_i} \le \max_i a_i + \log n$ [Boyd and Vandenberghe, 2006]. Equivalently, the min function is bounded by $\min_i a_i - \log n \le -\log \sum_i e^{-a_i} \le \min_i a_i$.

These identities allow us to prove two inequalities in Claim i, i.e. $\max_i \ell(z_i) = \ell(\min_i z_i) \le \ell\left(-\log \sum_i e^{-z_i}\right)$ and $\log \sum_i e^{\ell(z_i)} \le \max_i \ell(z_i) + \log n$. To prove the remaining inequality $\ell\left(-\log \sum_i e^{-z_i}\right) < \log \sum_i e^{\ell(z_i)}$, note that for the logistic loss $\ell\left(-\log \sum_i e^{-z_i}\right) = \log(1 + \sum_i e^{-z_i})$ and $\log \sum_i e^{\ell(z_i)} = \log(n + \sum_i e^{-z_i})$. Since $n > 1$, strict inequality holds.

To prove Claim ii, we need to show that $\ell\left(-\log \sum_i e^{-z_i}\right) = \log(1+\sum_i e^{-z_i}) < \sum_i \ell(z_i) = \sum_i \log(1 + e^{-z_i})$. This is equivalent to $1 + \sum_i e^{-z_i} < \prod_i (1 + e^{-z_i}) = \sum_{\mathbf{c} \in \{0,1\}^n} e^{-\mathbf{c}^{\mathrm{T}}\mathbf{z}} = 1 + \sum_i e^{-z_i} + \sum_{\mathbf{c} \in \{0,1\}^n, \mathbf{1}^{\mathrm{T}}\mathbf{c} > 1} e^{-\mathbf{c}^{\mathrm{T}}\mathbf{z}}$. Finally, we have $\sum_{\mathbf{c} \in \{0,1\}^n, \mathbf{1}^{\mathrm{T}}\mathbf{c} > 1} e^{-\mathbf{c}^{\mathrm{T}}\mathbf{z}} > 0$ because the exponential function is strictly positive.

To prove Claim iii, it suffices to find set of numbers $\{z_1, \ldots, z_n\}$ for which $\log \sum_i e^{\ell(z_i)} = \log(n + \sum_i e^{-z_i}) > \sum_i \ell(z_i) = \sum_i \log(1 + e^{-z_i})$. This is equivalent to $n + \sum_i e^{-z_i} > \prod_i (1 + e^{-z_i})$. By setting $(\forall i)\ z_i = \log n$, we reduce the claim we want to prove to $n + 1 > (1+\frac{1}{n})^n$. Strict inequality holds for $n > 1$. Furthermore, note that $\lim_{n \to +\infty} (1 + \frac{1}{n})^n = e$. □